STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 297–314, DOI 10.21307/stattrans-2018-017

A NEW METHOD FOR COVARIATE SELECTION IN COX MODEL

Ujjwal Das[1], **Nader Ebrahimi**[2]

ABSTRACT

In a wide spectrum of natural and social sciences, very often one encounters a large number of predictors for time to event data. An important task is to select right ones, and thereafter carry out the analysis. The ℓ_1 penalized regression, known as "least absolute shrinkage and selection operator" (LASSO) became a popular approach for predictor selection in last two decades. The LASSO regression involves a penalizing parameter (commonly denoted by λ) which controls the extent of penalty and hence plays a crucial role in identifying the right covariates. In this paper we propose an information theory-based method to determine the value of λ in association with the Cox proportional hazards model. Furthermore, an efficient algorithm is discussed in the same context. We demonstrate the usefulness of our method through an extensive simulation study. We compare the performance of our proposal with existing methods. Finally, the proposed method and the algorithm are illustrated using a real data set.

Key words: Bhattacharya distance, index of resolvability, Kullback-Leibler measure, ℓ_1 penalty, proportional hazards model, time to event data.

1. Introduction

The statistical analysis of time to event data is very common in several applied fields, such as biology, medicine, economics, engineering and social sciences. Typical examples of such an event may be the onset of a disease, death of a subject under study, occurrence of default of a corporate bond, malfunctioning of a system, etc. It is very frequent to adjust the analysis of those event times by incorporating the information available from covariates. One of the popular ways of analysing time to event data is based on the hazard rate function, and a common way of modelling the hazard rate function with covariate matrix Z is to write it as the product of the baseline hazard and some function of Z. This model referred to as 'proportional hazards' or the 'Cox model', can connect the covariates with time to event in a parametric or semi-parametric fashion. Mathematically, from Cox (1972) we have

$$h(t|Z) = h_0(t)\exp(Z'\beta), \tag{1.1}$$

where $h_0(t)$ is called the baseline hazard rate, $Z'\beta = \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p$ and $\exp(Z'\beta)$ describes how the hazard rate varies in response to covariates. One may

[1]Operations Management, Quantitative Methods and Information Systems Area, Indian Institute of Management, Udaipur 313001, Rajasthan, India. E-mail: ujjwal.das@iimu.ac.in.
[2]Division of Statistics, Northern Illinois University, Dekalb, IL 60115, USA.
E-mail: nader@math.niu.edu.

assume some parametric form for $h_0(t)$ and then (1.1) reduces to a parametric model. If no parametric form is assumed for $h_0(t)$ then the model (1.1) is semi-parametric. In practice, the estimation and inference from the Cox model is based on the partial likelihood function. But for our purpose we use the full likelihood function.

In some practical studies such as genetics, researchers may have a large number of covariates (p) from fewer number of observations n, and they may need to select only few of those many covariates. An example includes a typical microarray data set that consists of thousands of genes from hundred subjects. Traditional selection methods such as stepwise deletion or best subset selection though useful but may perform poorly in high dimensional $(p >> n)$ situations. The limitations of the existing methods of model selection are mentioned in Breiman (1996) and Fan and Li (2001). As a unified method of variable selection for both low and high dimension, the penalized approach has gained increasing popularity in recent years. The penalized methods with some conditions on the penalty functions, not only retain the good properties of the old methods but also enjoy theoretical justifications. Among the convex penalty functions, the least absolute shrinkage and selection operator or LASSO proposed by Tibshirani (1996) has gained enormous attention from the researchers. LASSO is defined as the ℓ_1 norm of the parameters: $\lambda||\beta||_1$, where β is the vector of regression coefficients and λ is the tuning parameter or penalizing parameter. The penalizing parameter plays an influential role for variable selection. A larger value of λ exerts a higher penalty on regression coefficients, resulting in inclusion of fewer variables in the model. Conversely, a small value of λ leads to less penalty, and hence inclusion of many variables. Commonly, a sequence of λ values is generated and then variables are detected for each value of the series. Thereafter, a value of λ is chosen by k-fold cross validation, and corresponding set of predictors are included in the model. Tibshirani (1997) used generalized cross validation for the Cox model. More recently, Simon et al. (2011) developed an R-package for variable selection in Cox model via LASSO with λ selected thorough cross-validation. Li and Barron (2000) developed the concept of information theoretically valid ℓ_1-penalty by extending the work of Grunwald (2007). Using a similar risk analysis Barron *et al.* $(2008a)$ and, Barron and Luo (2008) developed the concept of information theoretically valid ℓ_1 norm penalty function for linear models. They obtained a lower bound on the penalizing parameter which makes the LASSO penalty information theoretically valid. Recently, Das and Ebraimi (2017) extended the concept for accelerated failure time model. In this paper, we introduce the information theory for time to event data under the model (1.1) and obtain the bound for λ. The nonlinear structure of the model (1.1) makes the results more intricate than linear models. We will use the lower bound as the value of the initial penalizing parameter. In addition to that, we propose an efficient algorithm for the Cox proportional hazards model for variable selection following Barron et al. (2008). Any software that performs constrained optimization, can be used to implement the proposed algorithm.

The paper is organized as follows. A brief description on information theory

along with related concepts and, the determination of the bound on penalizing parameter for the Cox model are given in subsections 2.1 and 2.2, respectively. Section 3 deals with the algorithm and its accuracy. Section 4 ensures the usefulness of the proposed methodology through extensive simulation studies. The results are presented in a tabular format for different combinations of n and p with different censoring proportions. The performance of our method is compared with existing methods of selecting the tuning parameter in the immediate section. In Section 6 we illustrate our proposed method using a real world data, and compare the results with other methods. Finally, some concluding remarks complete the paper in Section 7.

2. Method

2.1. Preliminaries

This section provides a brief summary of different information measures. For a detailed discussion one can see Ebrahimi *et al.* (2010). The most well-known and widely used measure of uncertainty is Shannon's entropy (Shannon, 1948). For a random variable X with a domain S, its entropy $H(X)$ is defined as $-\int_S \log p(x) dP(x)$, where $P(x)$ is the cumulative distribution function and $p(x)$ is the probability density (mass) function of X. As a measure of information discrepancy between two probability distribution functions P and Q, we use Kullback-Leibler (KL) divergence (Kullback, 1959) given by $D(P,Q) = E_p \log(\frac{p}{q}) = \int_S \log(\frac{p(x)}{q(x)}) dP(x)$, provided P is absolutely continuous with respect to Q on the support S. Bhattacharya distance is an alternative way to discriminate between two distribution functions P and Q, and it is given by

$$d(P,Q) = -2\log \int \sqrt{p(x)q(x)}dx, \tag{2.1}$$

see Bhattacharya, (1943). Throughout this paper, Bhattacharya distance is used as the loss function to judge the accuracy of the estimate.

Index of Resolvability: Let L_f be the likelihood characterized by f and f^* is the true value of f. Then, the index of resolvability is defined as

$$R_n(f^*) = \min_{f \in \mathscr{F}} \left\{ \frac{1}{n} D(L_{f^*}, L_f) + \frac{1}{n} pen(f) \right\}, \tag{2.2}$$

where f is a candidate to estimate unknown f^*, $\mathscr{F}$ is the set of all possible values of f and $pen(f)$ denotes some penalty function. We use this index to upper-bound the statistical risk, associated with the estimates obtained by achieving the following minimization

$$\min_{f \in \mathscr{F}} \left\{ \frac{1}{n} \log(\frac{1}{L_f}) + \frac{1}{n} pen(f) \right\}. \tag{2.3}$$

The estimator obtained from (2.3) is called minimal complexity estimator. It can be shown that the expression under minimization in (2.3), converges in probability to

index of resolvability plus a constant (entropy), which ensures that the minimization in (2.3) is equivalent to the minimization of the resolvability index, $R_n(f^*)$ in (2.2). For more details see Barron et al. (2008).

From (1.1) f^* is the linear predictor given by $Z'\beta^*$. Let $\hat{f}$ be the minimal complexity estimator of f^*. Then we measure the associated risk of $\hat{f}$ by $E[d(L_{f^*}, L_{\hat{f}})]$. We choose the penalizing parameter of LASSO such that

$$E\left(\bar{d}(L_{f^*}, L_{\hat{f}})\right) \leq \inf_{\beta \in \mathscr{R}^p} \{\bar{D}(L_{f^*}, L_f) + \frac{\lambda}{n} \sum_{j=1}^{p} |\beta_j|\}. \tag{2.4}$$

where $\bar{d}(L_{f^*}, L_{\hat{f}}) = d(L_{f^*}, L_{\hat{f}})/n$ and $\bar{D}(L_{f^*}, L_f) = D(L_{f^*}, L_f)/n$ are the average Bhattacharya distance and Kullback-Leibler measure respectively, when averaged across the n independent subjects. In the next subsection we provide a lower bound of λ so that the risk bound in (2.4) holds for Cox model.

2.2. Determination of the bound on penalizing parameter

We consider survival studies in which n individuals are put on test and data of the form (v_i, δ_i, z_i) for $i = 1, 2, ..., n$, are collected. Here, v_i is the minimum of the exact failure time X_i and the censoring time C_i of the i^{th} individual, $\delta_i = I(X_i \leq C_i)$ is an indicator variable that represents the failure status, and z_i is the corresponding covariate that may be a vector. In addition, the survival function of the i^{th} individual is $S(t|z_i) = P(X_i > t|z_i)$. The corresponding density function is $f(t|z_i)$, where X_i is the exact failure time. Furthermore, we assume that the censoring time C_i of the i^{th} individual is a random variable with survival and density functions $G(t|z_i)$ and $g(t|z_i)$ respectively, and that given $z_1, ..., z_n$, the $C_1, ..., C_n$ are stochastically independent of each other and of the independent failure times $X_1, ..., X_n$. Therefore, the full likelihood function of the data (t_i, δ_i, z_i), conditional on $z_1, ..., z_n$, is

$$L_f(v_1, v_2, ..., v_n|\delta_1, \delta_2, ..., \delta_n, Z) = \prod_{i=1}^{n} (f(x_i|Z_i)G(x_i|Z_i))^{\delta_i} (S(C_i|Z_i)g(C_i|Z_i))^{1-\delta_i}$$

Since the censoring time is noninformative, the full likelihood function can be rewritten as

$$\begin{aligned} L_f(v_1, v_2, ..., v_n|\delta_1, \delta_2, ..., \delta_n, Z) &\propto \prod_{i=1}^{n} (f(x_i|Z_i))^{\delta_i} (S(C_i|Z_i))^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(h_0(x_i) \exp[-H_0(x_i) \exp(f_i)]\, e^{f_i}\right)^{\delta_i} \\ &\quad (\exp[-H_0(C_i) \exp(f_i)])^{1-\delta_i} \end{aligned} \tag{2.5}$$

Under the above likelihood we have the following bound on λ:
Result 1: The ℓ_1 penalized likelihood estimator $\hat{f} = f_{\hat{\beta}} = Z_i'\hat{\beta}$ obtained by

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \frac{\delta_i}{n} \left[H_0(x_i) e^{Z_i'\beta} - Z_i'\beta \right] + \sum_{i=1}^{n} \left[(1-\delta_i) \frac{H_0(C_i)}{n} e^{Z_i'\beta} \right] + \frac{\lambda}{n} ||\beta||_1 \right\} \tag{2.6}$$

attains the risk bound

$$E\bar{d}(L_{f^*}, L_{\hat{f}}) \leq \inf_{\beta} \left\{ \bar{D}(L_{f^*}, L_f) + \frac{\lambda}{n} \sum_{j=1}^{p} |\beta_j| \right\}$$

for every sample size provided that

$$\lambda \quad \geq \quad 2\sqrt{\log 2p} \sqrt{\left[\sum_{i=1}^{n} \left\{ \delta_i \left(H_0(x_i) e^{f_i} - \frac{1}{2} \right) + (1-\delta_i) \left(H_0(C_i) e^{f_i} - \frac{1}{2} \right) \right\} \right]}. \tag{2.7}$$

In practice, f_i is replaced by $\hat{f}_i$ obtained from (2.6).
Proof: The proof is outlined in the Appendix.
Remark: Under certain conditions (2.7) may not work. In that case, the bound will be

$$\lambda \quad \geq \quad 2\sqrt{\log 2p} \sqrt{\sum_{i=1}^{n} [\delta_i e^{f_i} H_0(x_i) + (1-\delta_i) e^{f_i} H_0(C_i)]}. \tag{2.8}$$

The condition is discussed in the Appendix.

3. The Algorithm

We propose an algorithm for the detection of regression parameters in the Cox model following Barron et al. (2008). For $(p < n)$ we fit the Cox-model to the data and use the point estimates as initial estimate for the algorithm. For $(p > n)$ we begin with $\beta^0 = \mathbf{0}$. Then, we estimate the cumulative baseline hazard by using the Breslow-type estimator. With these, next we estimate λ by using (2.7) or (2.8) according to the necessity. For any $t \geq 1$ we will move from $(t-1)^{th}$ step to t^{th} step of iteration by: $\beta^t = \alpha\beta^{t-1} + \gamma I_l$, where the parameters: $\alpha \in [0,1]$, $\gamma \in (-\infty, \infty)$, and I_l is a vector of zero except for l^{th} component which is 1. Combining this with (2.6)

the likelihood, as a function of α and γ, becomes

$$\begin{aligned} W^t(\alpha,\gamma,l) &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[H_0(x_i)\exp\left(\alpha\sum_{j=1}^{p}Z_i'\beta^{t-1}+\gamma Z_{il}\right)-\left(\alpha\sum_{j=1}^{p}Z_i'\beta^{t-1}+\gamma Z_{il}\right)\right] \\ &\quad+\frac{1}{n}\sum_{i=1}^{n}\left[(1-\delta_i)H_0(C_i)\exp\left(\alpha\sum_{j=1}^{p}Z_i'\beta^{t-1}+\gamma Z_{il}\right)\right] \\ &\quad+\frac{\lambda}{n}\left(\alpha\sum_{j=1}^{p}|\beta_j^{t-1}|+|\gamma|\right). \end{aligned} \tag{3.1}$$

for every coordinate $l=1,2,...,p$. Now, we minimize (3.1) with respect to α and γ and obtain the value of the objective function for each $l=1,2,...,p$. At t^{th} iteration the optimal α_t, γ_t and $I_{l(t)}$ are those for which the value of the objective function is minimum. We change that coordinate(s) and set others to zero. At the end of each iteration the estimates of λ and cumulative baseline hazards are also updated for the next iteration. The process is repeated until no new covariate is detected and the absolute difference between the estimates from two consecutive iterations is less than some preassigned small number. Any standard software can be used for performing the constrained optimization. We call R-routine *'constrOptim'* with the option *'Nelder-Mead'* method for its suitability to optimization of non-smooth functions. The R-code can be available from the corresponding author upon request.

3.1. Convergence of the Algorithm

Let L_f be the likelihood function with unknown parameters (or linear combination of parameters) f as given in (2.5), estimated by $\hat{f}_{(k)}$ at k^{th} iteration. Then we have
Result 2: Let $L_{\hat{f}}$ be the minimal complexity estimate of L_{f^*} and $L_{\hat{f}_k}$ be the estimate from k^{th} iteration obtained by our proposed algorithm. Then,

$$\frac{1}{n}\log\frac{1}{L_{\hat{f}_{(k)}}(x)}+\lambda u_{(k)}\leq\inf_f\left\{\frac{1}{n}\log\frac{1}{L_f(x)}+\lambda U_f+\frac{4U_f^2}{k+1}\right\}, \tag{3.2}$$

where $u_{(k)}=\sum_{j=1}^{p}|\hat{\beta}_{j,(k)}|$ and $U_f=\sum_{j=1}^{p}|\beta_j|$ with $\hat{\beta}_{j,(k)}$ is the estimate of β_j at k^{th} iteration.
Proof: The proof is given in the Appendix.

4. Numerical Studies

We investigate the performance of the proposed λ along with the algorithm through simulations. We will use the lower bound of λ as its value, for all numerical investigations. First, we create a matrix of 100 rows and 1000 columns by randomly drawing 1000 observations from a 100-dimensional multivariate normal distribution with mean **0** and pairwise correlation 0.1. Throughout the simulation study, we keep

this matrix fixed and use appropriate number of rows and columns as design matrix under four different scenarios: (a) n=100, p=50, (b) n=1000, p=100 for low dimension, and (c) n=50, p=100, (d) n=100, p=1000 for high dimension. For (a) we use the first 50 columns of the matrix, for (b) we transpose the matrix, and for (c) we consider the first 100 columns with their first 50 rows for numerical studies. Let β denote the true vector of regression coefficients. So, β is a vector of length 50 for (a), of length 100 for (b) and (c), and of length 1000 for (d). In each case, we randomly choose seven elements of β and set them to unity, and rest of the elements are all zero. Let Z be the design matrix of appropriate order. We model the baseline hazard of Cox regression assuming Weibull distribution to generate data. In this way we get a closed form expression for the survival function. We equate the survival function with random numbers generated from uniform(0,1) distribution, and then invert the survival function to get the time to event. We choose the scale and shape parameters of the Weibull distribution as 1 and 1.2 respectively. For a detailed discussion on the methods to generate data from the Cox model, one can see Bender, Augustin and Blettner (2005). Except for (n=50, p=100), for remaining pairs of (n, p) we vary the censoring proportion from 5% to 40$\%$ with an increment of 5%. In this way, we generate 1000 data sets for every combination of n, p and censoring percentage. Before analysis, for all the eight covariates we subtract the respective mean and then divide them by the respective standard deviation. Then, the variables are selected through the algorithm discussed in Section 3. The simulation results are summarized in Table 1, where **n** represents the number of subjects, **p** is the number of covariates as candidate of the model, **Cens. Pcnt.** gives percent of censoring, **TMDR** is the true model detection rate defined as the percentage of replications where the full model (all correct seven covariates) is detected, **Median** and **Mean** the number of correct variables detected, and **Avg. Incln.** is the average model size, from the 1000 replications.

From Table 1 we find that the method is working well for detecting the correct set of variables except for $n = 50, p = 100$. Along with the median, the average number of correct variables included is also higher than 6 for both $n = 100, p = 50$ and $n = 100, p = 1000$. We note that the average model size is not far from the average number of correct covariates detected, in all cases considered here. The phenomena indicates the inclusion of fewer false variables. More specifically, for $n = 1000, p = 100$, the entire correct model is identified always without any error for all censoring percentages. We observe that the convergence was achieved equally faster whether the initial estimate was $\mathbf{0}$ or taken from the Cox model fitting, for low dimension.

5. Comparison

We compare our proposed method of tuning parameter selection with cross-validation (CV), generalized cross-validation (GCV) and BIC. We use **R**-package *glmnet*. For a detailed discussion on the *glmnet*, its algorithm and convergence, see Simon *et al.* (2011). We reconsider the simulated data sets from Section 4, and reanalyse them

Table 1: Summary of the Simulation Studies

n	p	Cens. Pcnt.	TMDR	Median	Mean	Avg. Incln.
50	100	5.71	15.01	5	5.39	5.81
		10.88	10.1	5	4.89	5.34
		14.19	6.6	5	4.66	5.02
		19.68	1.1	4	4.12	4.66
		25.31	0	4	3.78	4.73
		29.19	0	4	3.72	4.81
100	50	5.27	90.1	7	6.87	7.05
		10.16	84.1	7	6.75	7.06
		15.22	80.8	7	6.71	7.1
		20.27	76.7	7	6.65	7.08
		25.02	74.9	7	6.64	7.09
		31.2	72.8	7	6.62	7.15
		34.87	68.1	7	6.54	7.06
		39.42	66.5	7	6.52	6.83
100	1000	5.76	84.4	7	6.79	7.58
		10.61	80.9	7	6.76	7.62
		14.85	77.2	7	6.74	7.53
		20.25	74.3	7	6.72	7.12
		25.25	71.8	7	6.65	7.24
		30.11	69.4	7	6.59	7.19
		35.71	65.9	7	6.57	7.22
		41.48	62.4	7	6.44	7.61
1000	100	5.02	100	7	7	7
		10.11	100	7	7	7
		15.01	100	7	7	7
		20.09	100	7	7	7
		25.19	100	7	7	7
		30.15	100	7	7	7
		34.9	100	7	7	7
		40.11	100	7	7	7

by *glmnet* in conjunction with *10-fold* CV, GCV and BIC. *10-fold* CV was performed using the R-function *cv.glmnet*, when for the other two we fit the Cox model with the selected predictors for every value of λ and then obtain the GCV and BIC values for each model. From a sequence of λ values, we pick the one as the value of the penalizing parameter and the corresponding model, for which the desired criterion (CV, GCV or BIC) attains its minimum. We compare the average number of variables detected for different values of n, p and censoring percentages, from all the methods. Table 2 provides the average number of predictors identified as non-zero from 1000 replications for 5% to 40% censoring. For $n = 50$ and $p = 100$ we perform the simulation up to 30% censoring. From Table 2 we see that cross-validation

Table 2: Average number of variables detected by Our Method and cross-validation

n	p	Method	5%	10%	15%	20%	25%	30%	35%	40%
50	100	Our method	5.81	5.34	5.02	4.66	4.73	4.81		
		CV	15.51	15.87	15.93	15.27	15.43	15.14		
		GCV	15.56	15.25	16.09	15.36	15.34	15.39		
		BIC	14.15	13.98	14.53	14.07	14.14	14.01		
100	50	Our method	7.05	7.06	7.1	7.08	7.09	7.15	7.06	6.83
		CV	20.15	19.63	19.92	18.99	20.19	18.75	20.19	19.09
		GCV	20.57	20.08	19.93	19.69	20.74	20.14	21.01	20.68
		BIC	15.21	15.17	14.94	14.61	15.33	15.18	15.66	14.88
100	1000	Our method	7.58	7.62	7.53	7.12	7.24	7.19	7.22	7.61
		CV	29.39	30.12	30.67	31.76	30.51	30.32	30.48	30.47
		GCV	30.15	31.87	31.29	32.27	31.52	31.11	31.61	31.83
		BIC	24.25	25.07	24.79	24.82	25.33	25.11	25.02	24.90
1000	100	Our method	7	7	7	7	7	7	7	7
		CV	38.93	37.71	38.13	37.82	36.57	36.93	37.12	36.67
		GCV	39.51	38.96	39.09	38.76	38.03	37.84	37.94	38.02
		BIC	30.01	29.71	29.37	29.58	29.12	29.17	29.93	29.66

tends to select more covariates compared to our method. For both $n = 100, p = 50$ and $n = 100, p = 1000$ the average model size by our method is near 7, whereas from cross-validation and GCV these are around 19 and 30 respectively. Similarly, for $n = 1000$ and $p = 100$ our method detects all the covariates up to 40% censoring without any false inclusion whereas the average model size from cross-validations is more than 35. The BIC tends to select fewer variables than CV and GCV but higher than our proposed method. We note that *glmnet* is almost always able to identify the correct set of covariates for the simulated data sets. For example, for $n = 50$, $p = 100$ and with 30% censoring, the true model detection rate (**TMDR**) was higher than 95% when our proposed method was unable to find all the correct covariates in a single instance. So, the cross-validations and BIC tend to select an entire set of right predictors at the cost of larger model size. Additionally, the coordinate descent algorithm seems to be faster than our algorithm. In general, we see that the proposed method may not always detect the full model, but the inclusion of a false covariate is small compared to the cross-validations and BIC, for all the scenarios we considered here.

6. Real data analysis

We analyse data on survival of the patients with advanced lung cancer. The study was conducted by North Central cancer treatment group, and described in Loprinzi et al. (1994). After some cleaning we are left with survival time on 167 subjects with information on 8 covariates. The covariates are: institution code, age in years, gender, ECOG performance score, Karnofsky performance score rated by physicians, Karnofsky performance score rated by the patients, calories consumed at meals, and weight loss in last six months. We analyse the data in three different ways, and as before, calculate the Bayesian information criterion (BIC) of the final selected model from each method for comparison. First, we select the model by BIC. The resulting model includes only two covariates: gender and ECOG performance score. BIC of this model is 1006.99, when the BIC value of the full model is 1023.48. Next, we fit the ℓ_1-penalized Cox proportional hazard model with penalizing

parameter chosen by 10-fold CV, GCV and BIC by using R-package *glmnet*. The CV and GCV identify seven out of eight covariates (exclude the variable calories consumed), and BIC after fitting the Cox model with these seven selected covariates is 1018.36. When the penalizing parameter was selected through BIC, two more predictors (age and Karnofsky performance score rated by the patients) are dropped from the model. BIC of the Cox model with these five predictors is 1011.38. Finally, we analyse the data by our proposed method. As mentioned before, we use the lower bound from (2.7) as the value of λ. Our method detects three covariates: institution code, gender and ECOG performance score. The BIC of the Cox model with these three variables comes out as 1008.66. We note that the p-values of gender and ECOG performance score are significant at 5% level when the same for the institution code (p-value$= 0.0675$) is significant at 10% level. The result seems to be consistent with our finding in Section 5.

7. Conclusion

The selection of appropriate penalty parameter has great influence on variable selection. Cross-validation is a widely used approach for choosing the parameter. Leng, Lin and Wahba (2006) suggested to go with some method other than cross-validations or BIC when covariate selection is of primary importance. The numerical results show that the model resulted from our method always includes fewer non-active covariates. From that perspective our method may be thought of as an alternative route to choose the penalizing parameter. Certainly, the proposed method is not a panacea for variable selection when event time is the outcome of interest. We have seen in Sections 2 and 7 that for low dimension our method yields the model with second smallest BIC value. But BIC-based model selection cannot be performed in high dimension where penalized regression is the only tool for variable selection. In general, for many of the situations we study in this paper, our method shows promising results. Together with these, our method may be a good candidate when covariate selection is the primary goal.

REFERENCES

BARRON, A. R., COHEN, A., DAHMEN, W., DEVORE, R., (2008). Approximations and learning by greedy algorithms, The Annals of Statistics. 36, pp. 64–94.

BARRON, A. R., HUANG, C., LI, J. Q., LUO, X., (2008a). The MDL principle, penalized likelihoods, and statistical risk, In Festschrift for Jorma Rissanen, Tampere University Press.

BARRON, A. R., LUO, X., (2008). MDL Procedures with ℓ_1 Penalty and their Statistical Risk, Proceedings Workshop on Information Theoretic Methods in Science and Engineering.

BENDER, R., AUGUSTIN, T., BLETTNER, M., (2005). Generating survival times to simulate Cox proportional hazards models, Statistics in Medicine, 24 (11), pp. 1713–1723.

BHATTACHARYA, A., (1943). On a measure of divergence between two statistical populations defined by probability distributions, Bulletin of Calcutta Mathematical Society. 35, pp. 99–109.

BREIMAN, L., (1996). Heuristics of instability and stabilization in model selection, The Annals of Statistics. 24, pp. 2350–2383.

COX, D. R., (1972). Regression models and life tables (with discussions), Journal of Royal Statistical Society, Series B. 34, pp. 187–220.

DAS, U., EBRAHIMI, N., (2017). Covariate selection for accelerated failure time data, Communications in Statistics: Theory and Methods. 46, pp. 4051–64.

EBRAHIMI, N., SOOFI, E. S., SOYER, R., (2010). Information measures in perspective, International Statistical Review. 78, pp. 383–412.

FAN, J.„ LI, R., (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of American Statistical Association. 456, pp. 1348-1360.

GRUNWALD, P., (2007). The minimum description length principle. MIT Press, Cambridge, MA.

KULLBACK, S., (1959). Information theory and Statistics. Wiley, New York; (reprinted in 1968 by Dover).

LENG, C., LIN, Y., WAHBA, G., (2006). A note on the lasso and related procedures in model selection, Statistica Sinica, 16, pp. 1273–1284.

LI, J. Q., BARRON, A. R., (2000). Mixture density estimation, S Solla, T Leen and K.R. Muller (Eds.), Advances in Neural Processing Information System, 12, pp. 279–285.

LOPRINZI, C. L., LAURIE, J. A., WIEAND, H. S., KROOK, J. E., NOVOTNY, P. J., KUGLER, J. W., BARTEL, J., LAW, M., BATEMAN, M., KLATT, N. E. et al., (1994). Prospective evaluation of prognostic variables from patient-completed

questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology, 12 (3), pp. 601–607.

LUO, X., (2009). Penalized Likelihoods: Fast algorithms and risk bounds, Ph.D. Thesis, Statistics Department, Yale University.

ROSSI, P. H., BERK, R. A., LENIHAN., K. J., (1980). Money, Work and Crime: Some Experimental Results. New York: Academic Press.

SHANNON, C. E., (1948). A mathematical theory of communication, Bell Sys. Tech. J.. 27, pp. 379–423.

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), pp. 1–13.

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

Tibshirani, R., (1997). The lasso method for variable selection in the Cox model, Statistics in Medicine, 16, pp. 385–395.

APPENDIX

Here, we outline the proof of Result 1 and show the convergence of the proposed algorithm.

Proof of Result 1:

From Barron et al. (2008) the condition on penalty function is

$$pen(f) \geq \log(\frac{L_f(X)}{L_{\tilde{f}}(X)}) - 2\log \frac{E\sqrt{\frac{L_f(X)}{L_{f^*}(X)}}}{E\sqrt{\frac{L_{\tilde{f}}(X)}{L_{f^*}(X)}}} + 2\mathscr{L}(\tilde{f}), \tag{7.1}$$

Using the full likelihood and the fact that $\tilde{f}_i \xrightarrow{p} f_i$ we get,

$$\begin{aligned}\frac{L_f(v_1,v_2,...,v_n)}{L_{\tilde{f}}(v_1,v_2,...,v_n)} &= \prod_{i=1}^{n} \frac{\left(h_0(v_i)\exp\left[-H_0(v_i)e^{f_i}\right]e^{f_i}\right)^{\delta_i}\left(\exp\left[-H_0(C_i)e^{f_i}\right]\right)^{1-\delta_i}}{\left(h_0(v_i)\exp\left[-H_0(v_i)e^{\tilde{f}_i}\right]e^{\tilde{f}_i}\right)^{\delta_i}\left(\exp\left[-H_0(C_i)e^{\tilde{f}_i}\right]\right)^{1-\delta_i}} \\ &= \prod_{i=1}^{n}\left(\frac{\exp\left[-H_0(v_i)e^{f_i}\right]e^{f_i}}{\exp\left[-H_0(v_i)e^{\tilde{f}_i}\right]e^{\tilde{f}_i}}\right)^{\delta_i}\left(\frac{\exp\left[-H_0(C_i)e^{f_i}\right]}{\exp\left[-H_0(C_i)e^{\tilde{f}_i}\right]}\right)^{1-\delta_i}.\end{aligned}$$

Thus, by Taylor expansion up-to order 2, we have

$$\begin{aligned}\log\left(\frac{L_f(v_1,v_2,...,v_n)}{L_{\tilde{f}}(v_1,v_2,...,v_n)}\right) &= \sum_{i=1}^{n}[\delta_i\left\{H_0(v_i)(e^{\tilde{f}_i}-e^{f_i})+(f_i-\tilde{f}_i)\right\}+(1-\delta_i)H_0(C_i) \\ &\quad (e^{\tilde{f}_i}-e^{f_i})] \\ &= \sum_{i=1}^{n}\left[\delta_i H_0(v_i)e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}+(1-\delta_i)H_0(C_i)e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\right] \\ &= \sum_{i=1}^{n}e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\left[\delta_i H_0(v_i)+(1-\delta_i)H_0(C_i)\right],\end{aligned}$$

Next, consider the expectation from (7.1),

$$\begin{aligned}&E\sqrt{\frac{L_f(V_1,V_2,...,V_n)}{L_{f^*}(V_1,V_2,...,V_n)}} \\ &= \prod_{i=1}^{n}(\int_0^{C_i}\sqrt{\frac{\exp\left[-H_0(v_i)e^{f_i}\right]e^{f_i}}{\exp\left[-H_0(v_i)e^{f_i^*}\right]e^{f_i^*}}}h_0(v_i)\exp\left[-H_0(v_i)e^{f_i^*}\right]e^{f_i^*}dv_i + \\ &\sqrt{\frac{\exp\left[-H_0(C_i)e^{f_i}\right]}{\exp\left[-H_0(t_i)e^{f_i^*}\right]}}\exp\left[-H_0(C_i)e^{f_i^*}\right]) \\ = &\prod_{i=1}^{n}\left\{\int_0^{C_i}h_0(v_i)\exp\left[-\frac{H_0(v_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]e^{\frac{f_i+f_i^*}{2}}dv_i+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]\right\} \\ = &\prod_{i=1}^{n}\left\{\frac{2e^{\frac{f_i+f_i^*}{2}}}{e^{f_i}+e^{f_i^*}}\left[1-\exp\left(-H_0(C_i)\frac{e^{f_i}+e^{f_i^*}}{2}\right)\right]+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]\right\}.\end{aligned} \tag{7.2}$$

Hence, after some algebra the ratio of the expectation in (7.1) reduces to

$$\begin{aligned}
&\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}} \\
= &\prod_{i=1}^{n}\frac{\frac{2e^{\frac{f_i+f_i^*}{2}}}{e^{f_i}+e^{f_i^*}}\left[1-\exp\left(-H_0(C_i)\frac{e^{f_i}+e^{f_i^*}}{2}\right)\right]+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]}{\frac{2e^{\frac{\tilde{f}_i+f_i^*}{2}}}{e^{\tilde{f}_i}+e^{f_i^*}}\left[1-\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)\right]+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{\tilde{f}_i}+e^{f_i^*}\right)\right]} \\
= &\prod_{i=1}^{n}\left\{1+\frac{f_i-\tilde{f}_i}{2}+\frac{(f_i-\tilde{f}_i)^2}{8}\right\}\left(\frac{e^{f_i}+e^{f_i^*}+(\tilde{f}_i-f_i)e^{f_i}}{e^{f_i}+e^{f_i^*}}\right) \\
&\left\{\frac{1-\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)+(f_i-\tilde{f}_i)\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)H_0(C_i)\frac{e^{f_i}}{2}}{1-\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)}\right\} \\
= &\prod_{i=1}^{n}\left\{1+\frac{(f_i-\tilde{f}_i)^2}{8}\right\}. \qquad (7.3)
\end{aligned}$$

We expand f_i around $\tilde{f}_i$ up-to first order by Taylor series, and since $\tilde{f}_i \xrightarrow{p} f_i$ then by the fact that for x close to 0, $e^x = 1+x+\frac{x^2}{2}$. Taking $\log$ on both sides of (7.3) we get

$$\begin{aligned}
\log\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}} &= \sum_{i=1}^{n}\log\left\{1+\frac{(f_i-\tilde{f}_i)^2}{8}\right\} \\
&= \sum_{i=1}^{n}\frac{(f_i-\tilde{f}_i)^2}{8}. \qquad (7.4)
\end{aligned}$$

As a result, the second expression in (7.1) may be approximated as

$$-2\log\left(\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}\right) = -\sum_{i=1}^{n}\frac{(f_i-\tilde{f}_i)^2}{4}. \qquad (7.5)$$

Hence, together with (7.2) and (7.5) the condition (7.1) is equivalent to

$$\begin{aligned} pen(f) &\geq \sum_{i=1}^{n}\left[e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\{\delta_i H_0(v_i)+(1-\delta_i)H_0(C_i)\}-\frac{(\tilde{f}_i-f_i)^2}{4}\right]+2\mathscr{L}(\tilde{f}) \\ &= \sum_{i=1}^{n}\frac{(\tilde{f}_i-f_i)^2}{2}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]+2\mathscr{L}(\tilde{f}). \end{aligned} \tag{7.6}$$

Using the facts that $(\tilde{f}_i-f_i)^2 \xrightarrow{p} E(\tilde{f}_i-f_i)^2 = Var(\tilde{f}_i)$ and this variance has an upper bound $\frac{UU_f}{K}$ i.e. $Var(\tilde{f}_i) \leq \frac{UU_f}{K}$. This upper bound together with the fact that $\mathscr{L}(\tilde{f}) = K\log 2p$ yield an upper bound for the right-hand side of (7.6). Replacing these in (7.6) we obtain

$$pen(f) \geq \frac{UU_f}{2K}\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]+2K\log 2p. \tag{7.7}$$

Differentiating (7.7) with respect to K and then equating it to zero we get

$$K=\frac{\sqrt{UU_f}}{2\sqrt{\log 2p}}\sqrt{\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}.$$

Then, replacing that value of K in (7.7) with the choice of $U = U_f$ we get

$$\begin{aligned} pen(f) &\geq 2\sqrt{UU_f\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}\sqrt{\log 2p} \\ \Rightarrow \lambda U_f &\geq 2\sqrt{UU_f\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}\sqrt{\log 2p} \end{aligned}$$

which is equivalent to

$$\frac{\lambda}{n} \geq 2\frac{\sqrt{\log 2p}}{n}\sqrt{\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}. \tag{7.8}$$

This completes the proof of the theorem.
There is a chance that the sum in (7.7) can be negative. Then, we cannot proceed further with that negative sum. In that situation, we adopt a slightly modified route

to overcome this difficulty. From (7.5) we have

$$-2\log\left(\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}\right) \approx -\sum_{i=1}^{n}\frac{(f_i-\tilde{f}_i)^2}{4}\leq 0. \tag{7.9}$$

With the bound in (7.9) and by the facts that $(\tilde{f}_i-f_i)^2 \xrightarrow{p} E(\tilde{f}_i-f_i)^2 = Var(\tilde{f}_i) \leq \frac{UU_f}{K}$ and $\mathscr{L}(\tilde{f}) = K\log 2p$, the condition (7.1) reduces to

$$\begin{aligned} pen(f) &\geq \sum_{i=1}^{n}\left[e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\{\delta_i H_0(v_i)+(1-\delta_i)H_0(C_i)\}\right]+2\mathscr{L}(\tilde{f}) \\ &= \frac{UU_f}{2K}\sum_{i=1}^{n}\left[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)\right]+2K\log 2p. \end{aligned} \tag{7.10}$$

Then, we minimize the right-hand side of (7.10) with respect to K and choose $U = U_f$ as before. Now, the equation (7.10) reduces to

$$\begin{aligned} pen(f) &\geq 2\sqrt{UU_f\sum_{i=1}^{n}[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)]}\sqrt{\log 2p} \\ \Rightarrow \lambda U_f &\geq 2\sqrt{UU_f\sum_{i=1}^{n}[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)]}\sqrt{\log 2p} \\ \Rightarrow \frac{\lambda}{n} &\geq 2\frac{\sqrt{\log 2p}}{n}\sqrt{\sum_{i=1}^{n}[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)]}. \end{aligned} \tag{7.11}$$

From (7.11) it is clear that the penalty function is still information theoretically valid since it satisfies the condition (7.1).

Proof of Result 2:

Let $e_k = \frac{1}{n}\log\frac{L_{\hat{f}}(x)}{L_{\hat{f}_{(k)}}(x)} + \lambda(u_{(k)} - U_f)$. Then, using the full likelihood we get

$$\begin{aligned} e_k &= \frac{1}{n}\log\frac{L_{\hat{f}}(x)}{L_{\hat{f}_{(k)}}(x)} + \lambda(u_{(k)} - U_f) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left[\left(\frac{p_{\hat{f}}(x_i)}{p_{\hat{f}_{(k)}}(x_i)}\right)^{\delta_i}\left(\frac{\bar{P}_{\hat{f}}(C_i)}{\bar{P}_{\hat{f}_{(k)}}(C_i)}\right)^{1-\delta_i}\right] + \lambda(u_{(k)} - U_f) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left[\delta_i\left\{\hat{f}_i - \hat{f}_{i,(k)} + H_0(v_i)(e^{\hat{f}_{i,(k)}} - e^{\hat{f}_i})\right\} + (1-\delta_i)\log\frac{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)}\right] \\ &\quad + \lambda(u_{(k)} - U_f), \end{aligned} \tag{7.12}$$

where $\hat{f}_i = Z_i'\hat{\beta}$ and $\hat{f}_{i,(k)} = Z_i'\hat{\beta}_{(k)}$ with $\hat{\beta}_{(k)}$, obtained at k^{th} iteration, is the estimate of β. To prove the theorem we need to show that

$$e_k \leq (1-\alpha)e_{k-1} + \frac{1}{2}\alpha^2 U_f^2. \tag{7.13}$$

It is clear that to have the inequality (7.13), we only need to tackle the ratio of the survival functions from (7.12). For the i^{th} subject we rewrite the ratio of the survival functions from (7.12) in the following way

$$\begin{aligned} &\frac{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)} \\ &= \left[\frac{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)}\right]^{\bar{\alpha}} \frac{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\alpha}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\bar{\alpha}}}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)}. \end{aligned} \tag{7.14}$$

So, to prove (7.13) we need to show

$$\frac{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\alpha}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\bar{\alpha}}}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)} \leq 1. \tag{7.15}$$

Then, using the updating rule that $\hat{f}_{i,(k)} = \bar{\alpha}\hat{f}_{i,(k-1)} + \gamma Z_{il}$ we rewrite (7.15) in the following way

$$\begin{aligned} & \frac{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\alpha}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\bar{\alpha}}}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)} \\ = & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\bar{\alpha}}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\alpha}\exp\left(-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\gamma Z_{il}}\right)} \\ . & \end{aligned} \tag{7.16}$$

We choose customarily some α and $\gamma = \alpha U_f$ in such a way that $\gamma Z_{il} \xrightarrow{p} \alpha f_i$, which is estimated by $\alpha\hat{f}_i$. For more details regarding the customary choices and the convergence see [2]. Using these facts (7.16) reduces to

$$\begin{aligned} & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\bar{\alpha}}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\alpha}\exp\left(-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}\right)} \\ = & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\left\{\exp\left(-\bar{\alpha}H_0(C_i)e^{\hat{f}_i}\right)\right\}\left\{\exp\left(-\alpha H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}\exp\left(-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}\right)} \\ = & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\exp\left(-\bar{\alpha}H_0(C_i)e^{\hat{f}_i}-\alpha H_0(C_i)e^{\hat{f}_{i,(k-1)}}-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}\right)}. \end{aligned} \tag{7.17}$$

We denote the numerator and denominator of (7.17) by D_n and D_e, respectively. We see that for $\alpha = 0$ and 1, $D_n = D_e$ which reduces (7.17) to 1 and hence log of (7.17) becomes 0. For $\alpha \in (0,1)$ we study the nature of D_e. We have

$$\frac{\partial \log D_e}{\partial \alpha} = H_0(C_i)\left(e^{\hat{f}_i} - e^{\hat{f}_{i,(k-1)}} - e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}(\hat{f}_i - \hat{f}_{i,(k-1)})\right)$$

and

$$\frac{\partial^2 \log D_e}{\partial \alpha^2} = -H_0(C_i)\exp\left\{\bar{\alpha}\hat{f}_{i,(k-1)} + \alpha\hat{f}_i\right\}(\hat{f}_i - \hat{f}_{i,(k-1)})^2 \tag{7.18}$$

From (7.18) it is clear that $\log D_e$ and hence, D_e is strictly concave function. As a result, D_e cannot attain its maximum for $\alpha = 0$ or 1 since in that case D_e will be a constant. So, (7.17) is less than or equal to 1 for $0 < \alpha < 1$, indicating that its log is negative. This completes the proof of the result.