Applied Soft Computing 14 (2014) 498–503

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

The $L_{1/2}$ regularization method for variable selection in the Cox model

Cheng Liu [a], Yong Liang [a,*], Xin-Ze Luan [a], Kwong-Sak Leung [b], Tak-Ming Chan [b], Zong-Ben Xu [c], Hai Zhang [c]

[a] Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau
[b] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[c] Faculty of Science, Xi'an Jiaotong University, Xian, China

ARTICLE INFO

Article history:
Received 15 November 2011
Received in revised form 3 January 2013
Accepted 17 September 2013
Available online 19 October 2013

Keywords:
Survival analysis
Regularization
Variable selection
Cox model.

ABSTRACT

In this paper, we investigate to use the $L_{1/2}$ regularization method for variable selection based on the Cox's proportional hazards model. The $L_{1/2}$ regularization can be taken as a representative of L_q $(0<q<1)$ regularizations and has been demonstrated many attractive properties. To solve the $L_{1/2}$ penalized Cox model, we propose a coordinate descent algorithm with a new univariate half thresholding operator which is applicable to high-dimensional biological data. Simulation results based on standard artificial data show that the $L_{1/2}$ regularization method can be more accurate for variable selection than Lasso and SCAD methods. The results from real DNA microarray datasets indicate the $L_{1/2}$ regularization method performs competitively.

1. Introduction

Censored survival data analysis of DNA microarray data is an increasing interest in relating gene expression profiles to survival phenotypes such as time to cancer recurrence or death. The identification of gene signatures related to survival may provide new information and tools for clinical decision making, prognosis, diagnosis, choice of therapy, and may further aid in the search for new targets for drug design. The Cox proportional hazards model [1,2] is the most popular method for the censored survival data. However, due to the very high dimensional space of the predictors, i.e., the number of genes typically far exceeds sample size in the microarray experiments, the standard maximum Cox partial likelihood method cannot be applied directly to obtain the parameter estimates. To deal with the problem of high dimensionality, various approaches have proposed, for example, semi-supervised learning methods [3], supervised principal components [4], and the residual finesse approach [5].

Besides the high-dimensionality, the genes expression levels of some genes are often highly correlated, which creates the problem of high co-linearity. To deal with the problem of co-linearity, the most popular approach is to use the penalized partial likelihood, including both the L_2 penalized estimation [6], which is often called the ridge regression, and the L_1 penalized estimation [7], which is called the least absolute shrinkage and selection operator (Lasso) estimation. Comparing to the L_2 penalized procedure with constraints on the sum of the square of the coefficients, the Lasso procedure provides method for variable selection. However, the Lasso estimator does not possess the oracle properties, and several approaches beyond the lasso have been developed in recent years. Fan and Li [8] adopted the smoothly-clipped-absolute-deviation (SCAD) penalty which has better theoretical properties than the Lasso. Zhang and Lu [9] developed the adaptive Lasso method based on a penalized partial likelihood with adaptively weighted L_1 penalties on regression coefficients. Engler and Li [10] proposed the penalized Cox method with Elastic net penalty which is a linear combination of L_1 and L_2 penalties.

Inspired by the aforementioned methods, we proposed a new, efficient version of penalized Cox model, which based on a $L_{1/2}$ penalty. The $L_{1/2}$ penalty can be taken as a representative of L_q $(0<q<1)$ penalties in both sparsity and computational efficiency, and has demonstrated many attractive properties, such as unbiasedness, and oracle properties [11]. In this paper, we developed a coordinate descent algorithm with the $L_{1/2}$ regularization in the Cox model. The approach is applicable to high-dimensional data, such as DNA microarray datasets.

The rest of the paper is organized as follows. In Section 2, we proposed a new version of the penalized Cox model with the $L_{1/2}$ regularization. In Section 3, we developed the coordinate descent algorithm for the $L_{1/2}$ penalized Cox model. In Section 4, we evaluated the performance of our proposed approach on the simulated

* Corresponding author. Tel.: +853 88972034.
E-mail addresses: chengliu10@gmail.com (C. Liu), yliang@must.edu.mo (Y. Liang).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.09.006

and real DNA microarray datasets respectively. We concluded the article with Section 5.

2. $L_{1/2}$ penalized Cox model

Suppose the dataset has a sample size of n to study survival time T on covariate X, we use the data form of $(t_i, \delta_i, X_i)_{i=1}^n$ to represent the individual's sample, where t is the survival time, δ is the censoring indicator, if $\delta_i = 0$ indicates right censoring time and $\delta_i = 1$ indicates no censoring, $X_i = (x_{i1}, \ldots, x_{ip})$ represents the p-dimension covariates. By the Cox's proportional hazards model, the hazard function can be defined as: $h(t|\beta) = h_0(t)\exp(\beta^T X)$, where baseline hazard function $h_0(t)$ is unspecified or unknown and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the regression coefficient vector of p variables. Let ordered risk set at time $P(\beta) = \sum |\beta|^q$ be denoted by $R_r = \{j \in 1, \ldots, n : t_j > t_i\}$, based on the available sample data, to estimate the Cox model is to through minimizing the Cox's partial log likelihood function:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ x_i^T \beta - \log \left(\sum_{j \in R_i} \exp(x_j^T \beta) \right) \right\} \tag{1}$$

In practice, not all the n covariates may contribute to the prediction of survival outcomes: some components of $\beta_j (j = 1, 2, \ldots, p)$ may be zero in the true model. When the sample size goes to infinity, an ideal model selection and estimation procedure should be able to identify the true model with probability one, and provide consistent and efficient estimators for the relevant regression coefficients. Therefore, the regularization methods are applied to address this problem. When adding a regularization term (also call the penalty function) to Eq. (1), the penalized Cox model can be modeled as:

$$\beta = \text{argmin} \left\{ l(\beta) + \lambda \sum_{j=1}^{p} P(\beta_j) \right\} \tag{2}$$

where $\lambda > 0$ is a tuning parameter and $P(\beta)$ is the regularization term. Recently a series of regularization methods were proposed for the Cox's proportional hazards model, the popular regularization technique is Lasso (L_1) [12] which has the regularization term $P(\beta) = \sum |\beta|$. Many others L_1 type regularization terms also have been proposed to solve the Cox model, such as SCAD [8], adaptive Lasso [9]. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously.

The aforementioned penalized Cox model methods were based on the L_1 type penalties. Theoretically, the L_q $(0 < q < 1)$ type regularization $P(\beta) = \sum |\beta|^q$ with the lower value of q would lead to better solutions with more sparsity. However when q is very close to zero, difficulties with convergence arise. Therefore, Xu et al. [13] further explored the properties of the L_q $(0 < q < 1)$ regularizations and reveals the extreme importance and special role of the $L_{1/2}$ regularization. They proposed that when $1/2 < q < 1$, the $L_{1/2}$ regularization can yield most sparse results and its difficulty with convergence not very high compared with that of the L_1 regularization, while when $0 < q < 1/2$, the performance of the L_q penalties makes no significant difference and solving the $L_{1/2}$ regularization is much simpler than solving the L_0 regularization. Therefore, the $L_{1/2}$ regularization can be taken as a representative of the L_q $(0 < q < 1)$ regularizations. In this paper, we proposed an efficient version of the $L_{1/2}$ penalized Cox model as follows:

$$\beta_{1/2} = \text{argmin} \left\{ l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^{1/2} \right\} \tag{3}$$

The $L_{1/2}$ regularization has been demonstrated many attractive properties, such as unbiasedness, and oracle properties. The theoretical and experimental analyses show that the $L_{1/2}$ regularization is a competitive approach. Motivated by the fact, the penalized Cox model with the $L_{1/2}$ regularization is naturally expected.

3. A coordinated descent algorithm for $L_{1/2}$ penalized Cox model

The coordinate descent algorithm [14,15] is a "one-at-a-time" approach, which can apply to the L_1 type penalized regression. Its basic procedure of the coordinate descent algorithm is as follow: for each coefficient, to partially optimize the target function with respect to $\beta_j (j = 1, 2, \ldots, p)$, the remaining elements of β fixed at their most recently updated values, iteratively cycling through all coefficients until converged. Before introducing the coordinate descent algorithm for the $L_{1/2}$ penalized Cox model, we first consider the linear regression case. Suppose the dataset D has n samples, $D = (X_i, y_i)_{i=1}^n$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is ith input variables with dimensionality p and y_i is response variable for the ith observation. Assume the variables are standardized, so that $\sum_{i=1}^n x_{ij} = 0$, $\sum_{i=1}^n x_{ij}^2 = 1$ and $\sum_{i=1}^n y_i = 0$. The linear regression with the regularization term can be modeled as:

$$L(\beta) = \text{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - X'\beta)^2 + \lambda \sum_{j=1}^{p} P(\beta_j) \right\} \tag{4}$$

where $P(\beta)$ is the regularization term. The coordinate descent algorithm solves β_j and other $\beta_{k \neq j}$ are fixed ($k \neq j$ notes the parameters remained after jth element was removed). Eq. (4) can be rewritten as:

$$L(\beta) = \text{argmin} \left\{ \frac{1}{n} \left(y_i - \sum_{k \neq j} x_{ik}\beta_k + x_{ij}\beta_j \right)^2 + \lambda \sum_{k \neq j} P(\beta_k) + \lambda P(\beta_j) \right\} \tag{5}$$

The first order derivative at β_j can be estimated as:

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \left(-x_{ij} \left(y_i - \sum_{k \neq j} x_{ik}\beta_k - x_{ij}\beta_j \right) \right) + \lambda P\left(\beta_j\right)' = 0 \tag{6}$$

Define $\tilde{y}_i^{(j)} = \sum_{k \neq j} x_{ik}\beta_k$ as the partial residual for fitting β_j and $\omega_j = \sum_{i=1}^n x_{ij}(y_i - \tilde{y}_i^{(j)})$, the coordinate descent algorithm for the L_1 regularization (Lasso) can be solved by using the univariate soft thresholding operator:

$$\beta_j = S(\omega_j, \lambda) = \begin{cases} \omega_j + \lambda & \text{if } \omega_j < -\lambda \\ \omega_j - \lambda & \text{if } \omega_j > \lambda \\ 0 & \text{if } |\omega_j| < \lambda \end{cases} \tag{7}$$

Similarly, for the L_0 penalty, the thresholding operator of the coordinate descent algorithm can be defined as:

$$\beta_j = H(\omega_j, \lambda) = \omega I(|\omega_j| > \lambda) \tag{8}$$

$H(\omega_j, \lambda)$ is the univariate approximate solution to the L_0 regularization. It is equivalent to a hard thresholding operator.

According to Eqs. (7) and (8), we can know that the different penalties are associated with the different thresholding operators. Therefore, Xu et al. [13] proposed a half thresholding operator to solve the $L_{1/2}$ regularization for the linear regression model. It is an iterative algorithm and can be seen as a multivariate half thresholding approach. In this paper, a novel univariate half thresholding

operator of the coordinate descent algorithm for the $L_{1/2}$ regularization is proposed, and can be expressed as:

$$\beta_j = \text{Half}(\omega_j, \lambda) = \begin{cases} \frac{2}{3}\omega_j\left(1+\cos\left(\frac{2(\pi-\varphi_\lambda(\omega_j))}{3}\right)\right) & \text{if } |\omega_j| > \frac{3}{4}(\lambda)^{2/3} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Since the aforementioned coordinate descent algorithms cannot directly be applied for the nonlinear Cox model to obtain parameter estimates, Tibshirani [7] proposed a modified Newton–Raphson iterative procedure to reformulate the partial likelihood function of the penalized Cox model. Specifically, let $\eta = X\beta$, $\mu = -\partial l(\beta)/\partial\eta$, $A = -\partial^2 l/\partial\eta\eta^T$ and $z = \eta + A^{-}\mu$, by a one term Taylor expansion, $l(\beta)$ can then be represented as $(z-\eta)^T A(z-\eta)$. According to the approach of Gui and Li [16], this approximation can be rewritten as $(Y-\hat{X}\beta)^T(Y-\hat{X}\beta)$, where $\hat{z} = (A^{1/2})\cdot z$ and $\hat{X} = (A^{1/2})\cdot X$. Thus, we can directly apply the coordinate descent algorithm to the $L_{1/2}$ penalized Cox model with the Newton–Raphson iterative procedure, and the details are given follows:

The coordinate descent algorithm for the $L_{1/2}$ penalized Cox model
Step 1: Initial all $\beta_j = 0$ ($j = 1,2,\ldots,p$) and λ, set $m = 0$.
Step 2: Compute $\eta(m)$, $\mu(m)$, $A(m)$, $\hat{X}(m)$, $\hat{z}(m)$ based on current $\beta(m)$
Step 3: Solve $(\hat{z}(m)-\hat{X}\beta(m))^T(\hat{z}(m)-\hat{X}\beta(m)) + \lambda\sum_{j=1}^{p}\left|\beta_j(m)\right|^{1/2}$ repeat the following:
Cycle over $j = 1, \ldots, p$, until $\beta(m)$ not change:
Calculate $\omega_j(m) = \sum_{i=1}^{n}\hat{x}_{ij}(\hat{z}_i(m) - \hat{z}_i^{(j)}(m))$, where $\hat{x}_{ij}$ is a element of $\hat{X}(m) = (A^{1/2}(m))\cdot X$ and $\hat{z}_i^{(j)}(m) = \sum_{k\neq j}\hat{x}_{ik}\beta_k(m)$. Then update $\beta_j(m) = \text{Haft}(\omega_j, \lambda)$
Step 4: Let $m = m+1$, repeat Steps 2 and 3 until $\beta(m)$ convergence

Note that updating formulas by calculating weights samples $(\hat{X}(m) = (A^{1/2}(m))\cdot X)$ and pseudo-responses $(\hat{z}(m))$ based on the current $\beta(m)$ for each Step 2. The coordinate descent algorithm with the $L_{1/2}$ penalty works well in the sparsity problems, because the procedure does not need to change many irrelevant parameters and to recalculate partial residuals for each update step.

4. Simulation study

In order to assess performance of the $L_{1/2}$ penalized Cox model, several simulation studies were conducted under different data scenarios. The selection of the tuning parameters and the measures for comparing the performances of prediction models are introduced in Section 4.1. The comparison results of the $L_{1/2}$ penalized Cox model with other existing methods on the simulated and real DNA microarray datasets are reported in Sections 4.2 and 4.3, respectively.

4.1. Selection of tuning parameters and performance measures of prediction models

To select the tuning optimal parameter by cross-validation in the Cox model, we use a special approach proposed by van Houwelingen et al. [6], which is defined as:

$$\text{CV}(\lambda) = \sum_{i=1}^{k}\{l(\beta_{(-i)}(\lambda)) - l_{(-i)}(\beta_{(-i)}(\lambda))\} \quad (10)$$

where $\beta_{(-i)}(\lambda)$ represents the estimation of β that is obtained by the penalized method when the kth fold is left out. The term $l(\beta)$ is the log partial likelihood, and $l_{(-i)}(\beta)$ is the log partial likelihood without the ith fold. The optimal tuning parameter λ is obtained by maximizing CV(λ). Note that the choice of k will depend on the size of the dataset. In our experiments, we used 5-fold cross validation ($k = 5$).

The predictive performance measures of censored survival data are more complicated: these measures can only be computed if the case is not right censoring. Thus, several specially designed measure methods have been proposed in the literatures. In this paper, we employ the integrated Brier-Score (IBS) [17] and the concordance index (CI) [18] to evaluate the prediction ability of the survival data analysis methods.

4.1.1. Integrated Brier-Score (IBS)

The Brier Score (BS) is defined as a function of time $t > 0$ by:

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t|X_i)^2 1(t_i \le t \wedge \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t|X_i))^2 1(t_i > t)}{\hat{G}(t)}\right] \quad (11)$$

where $\hat{G}(\cdot)$ denotes the Kaplan–Meier estimation of the censoring distribution and $\hat{S}(\cdot|X_i)$ stands to estimate survival for the patient i. Note that the BS(t) is dependent on the time t, and its values are between 0 and 1. The good predictions at the time t result in small values of BS. The integrated Brier Score (IBS) is given by:

$$\text{IBS} = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} BS(t)\,dt \quad (12)$$

The IBS is used to assess the goodness of the predicted survival functions of all observations at every time between 0 and max(t_i).

4.1.2. Concordance Index (CI)

The Concordance Index (CI) can be interpreted as the fraction of all pairs of subjects which predicted survival times are correctly ordered among all subjects that can actually be ordered. By the CI definition, we can determine $t_i > t_j$ when $f_i > f_j$ and $\delta_j = 1$ where $f(\cdot)$ is survival function. The pairs for which neither $t_i > t_j$ nor $t_i < t_j$ can be determined are excluded from the calculation of CI. Thus, the CI is defined as:

$$\text{CI} = \frac{\sum_i\sum_j 1(f_i < f_j \wedge \delta_i = 1)}{\sum_i\sum_j 1(t_i < t_j \wedge \delta_i = 1)} \quad (13)$$

Note that the values of CI are between 0 and 1, the perfect predictions of the building model would lead to 1 while have a CI of 0.5 at random.

4.2. Analysis of simulated data

We carried out a simulation study to evaluate the parameter estimation, model selection and prediction capability performance of the three penalized Cox model methods, Lasso, SCAD and $L_{1/2}$ penalties. We adopted the Cox model simulation scheme in Bender's work [19]. The data generation procedure is as follows:

Step 1: We generate $\gamma_{i0}, \gamma_{i1}, \ldots, \gamma_{ip}$ ($i = 1,\ldots,n$) independently from standard normal distribution and set [20]: $X_{ij} = \gamma_{ij}\sqrt{1-\rho} + \gamma_{i0}\sqrt{\rho}$ ($j = 1,\ldots,p$) where ρ is the correlation coefficient.
Step 2: The survival time T_i ($i = 1,\ldots,n$) is constructed from a uniformly distributed variable U by $T_i = (1/\alpha)\log(1-(\alpha\times\log(U)/\omega\exp(\beta X)))$, where ω is the scale parameter, α is the shape parameter, β is the ground-true regression coefficients.
Step 3: Censoring time point T_i' ($i = 1,\ldots,n$, n indicates sample size) is obtained from an exponential distribution $E(\theta)$, where θ is determined by specify censoring rate.
Step 4: Here we define $t_i = \min(T_i, T_i')$ and $\delta_i = I(T_i \le T_i')$, the observed data represented as (t_i, δ_i, X_i) for the Cox model (Eq. (1)) are generated.

Table 1
Average number of genes and prognostic genes selected by the three methods on the simulated data in 50 runs.

Corr.	Size	Average of features selected			Average of true nonzero features			Recovery rate		
		Lasso	SCAD	$L_{1/2}$	Lasso	SCAD	$L_{1/2}$	Lasso	SCAD	$L_{1/2}$
$\rho = 0$	$n = 50$	29.1	7.7	4.7	1.6	1.54	1.44	0.054	0.201	0.327
	$n = 100$	38.4	7.2	6.1	2.78	2.72	2.76	0.072	0.377	0.452
	$n = 200$	45.2	10.3	9.3	2.96	2.89	2.98	0.065	0.311	0.320
$\rho = 0.3$	$n = 50$	30.2	6.8	5.2	1.64	1.54	1.38	0.054	0.226	0.260
	$n = 100$	36.8	7.4	6.2	2.80	2.64	2.64	0.076	0.356	0.420
	$n = 200$	50.7	12.1	11.5	2.96	2.94	2.94	0.058	0.242	0.255

Table 2
Average IBS and CI results of the Lasso, SCAD and $L_{1/2}$ penalized methods on the simulated datasets.

Corr.	Size	Average IBS			Average CI		
		Lasso	SCAD	$L_{1/2}$	Lasso	SCAD	$L_{1/2}$
$\rho = 0$	$n = 50$	0.1003	0.0997	0.1001	0.7031	0.7102	0.7044
	$n = 100$	0.0716	0.0716	0.0712	0.8406	0.8402	0.8431
	$n = 200$	0.0692	0.0697	0.0689	0.8970	0.8965	0.8974
$\rho = 0.3$	$n = 50$	0.1014	0.1041	0.1016	0.7309	0.7249	0.7320
	$n = 100$	0.0854	0.0874	0.0856	0.8032	0.8002	0.8036
	$n = 200$	0.0790	0.0783	0.0802	0.8572	0.8574	0.8551

For our experiments, we generate high-dimensional and low sample size datasets. In every simulation, the dimension of the predictor genes is $p = 500$, the coefficients of the three prognostic genes are: $\beta_1 = 3$, $\beta_4 = -2$, $\beta_7 = 1$, and the coefficients β of the remaining 497 genes are zeros. About 25% of the data are right censored. We consider the cases with the training sample size $n = 50, 100, 200$ and the correlation coefficients $\rho = 0$ and 0.3, respectively. To assess the variability of the experiment, each method is evaluated on a test set including 50 samples, and replicated over 50 random training and test partitions.

Table 1 shows the average number of genes and prognostic genes selected by three methods in 50 runs. The recovery rate is defined as the ratio of the average number of the prognostic genes to the average number of genes selected [20]. Overall the Lasso method selected the largest number of genes and the $L_{1/2}$ penalized method selected the least. When the training sample size is related small ($n = 50$), all the methods selected the prognostic genes difficultly. However, as n is increased, they can select almost all the three prognostic genes. The recovery rate of the $L_{1/2}$ penalized method was the highest, the SCAD was the second, and the Lasso was the lowest. All three methods perform better at the correlation coefficients $\rho = 0$ than at the correlation coefficients $\rho = 0.3$ as expected.

Fig. 1 displays the solution paths of the three methods in one selected simulation run. Here the x-axis displays the number of running steps, the y-axis is the coefficients. The optimal results obtained by the three methods are shown as vertical dotted lines. From Fig. 1, it clearly shows that the computational results of the $L_{1/2}$ penalized Cox model are more sparse than those of the Lasso and SCAD penalized methods. In summary, the $L_{1/2}$ penalized method can select almost all prognostic genes while discarding the vast of irrelevant genes.

To evaluate prediction performance of the three penalized Cox models, we presented their average IBC and CI values on the simulated datasets among 50 times in Table 2. In terms of IBC and CI, the three penalized methods performed similar for different parameter settings, and their differences seem to be marginal. Combined with the results reported in Table 1, we concluded that the $L_{1/2}$ penalized method showed similar predictive performance as the other methods with a more succinct model.

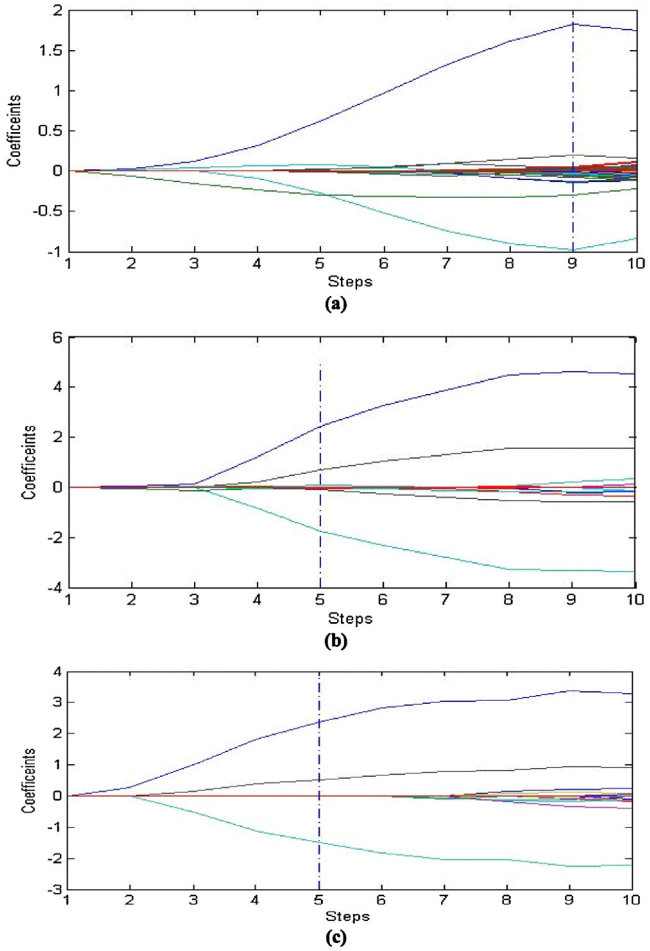

Fig. 1. The solution paths of the Lasso (a), SCAD (b) and $L_{1/2}$ (c) penalized methods in one sample simulation run, respectively.

Table 3
The detail information of 5 real gene expression datasets is used in the experiments.

Datasets	No. of genes	No. of samples	No. of censored	No. of training	No. of testing
DLBCL(2002)	7399	240	102	168	72
DLBCL(2003)	8810	92	28	64	28
DBC	4919	295	207	206	89
Lung cancer	7129	86	62	64	28
AML	6283	116	49	81	35

Table 4
Results of genes selected by the six methods on the 5 public datasets. In bold – the best performance.

Datasets	SPCR	PCR	PLS	Lasso	SCAD	$L_{1/2}$
DLBCL(2002)	3450	7399	7399	17.3	13.2	**10.5**
DLBCL(2003)	2387	8810	8810	10.1	4.2	**3.1**
DBC	2023	4919	4919	38.3	33.2	**23.6**
Lung cancer	1865	7129	7129	19.2	10.3	**8.4**
AML	3210	6283	6283	16.2	11.0	**9.8**
Mean	2587.0	6908.0	6908.0	20.2	14.3	11.1

Table 5
The IBS results obtained by the six methods on the 5 real datasets. In bold – the best performance.

Datasets	SPCR	PCR	PLS	Lasso	SCAD	$L_{1/2}$
DLBCL(2002)	**0.1970**	0.1974	0.2210	0.2074	0.2255	0.2232
DLBCL(2003)	0.1211	**0.1165**	0.1170	0.1203	0.1207	0.1205
DBC	0.1615	0.1625	0.1629	0.1561	0.1471	**0.1445**
Lung cancer	0.2034	0.1863	0.1953	0.1683	**0.1673**	0.1689
AML	0.2014	0.1932	0.1892	**0.1743**	**0.1743**	0.1781
Mean	0.1768	0.1772	0.1771	0.1653	0.1669	0.1670

4.3. Analysis of real microarray datasets

In this section, we evaluated the performance of the prediction methods on the real survival gene expression datasets. To further evaluate the performance of the $L_{1/2}$ penalized Cox model, we added three other type prediction methods which have been applied to high-dimensional biological data in this comparison: principal components regression (PCR) [21], supervised principal components regression (SPCR) [3,4], and partial least squares (PLS) [22].

Five publicly available datasets are used in this part. A brief description of these datasets is given below and summarized in Table 3.

Diffuse large B-cell lymphoma dataset (DLBCL)2002: This dataset published by Rosenwald et al. [23]. The dataset consists of 240 samples from patients. For each sample, 7399 gene expression measurements were obtained. The clinical outcome was survival time, either observed or censored.

Diffuse large B-cell lymphoma dataset (DLBCL)2003: This dataset is from Rosenwald et al. [24]. It consists of 92 lymphoma patients, and each patient contain 8810 genes.

Dutch breast cancer dataset (DBC): The Dutch breast cancer dataset is from van Houwelingen et al. [6], and consists of survival times and gene expression measurements from 295 women with breast cancer. The expression levels of $p = 4919$ genes were available.

Lung cancer dataset: The lung cancer dataset is from Beer et al. [25]. It consists of gene expressions of 4966 genes for 83 patients. The survival time as well as the censoring status is available.

AML dataset: The AML dataset is from Bullinger et al. [26]. It contains the expression profiles of 6283 genes for 116 patients, and the number of censored cases is 49.

We evaluated the prediction accuracy of the six estimated models using random partition: the training set of about 2/3 of the patients used for estimation and the test set of about 1/3 of the patients used for testing of the prediction capability. For estimating λ, we employed the 5-fold cross validation scheme using the training set. We repeated each procedure over 50 times.

Table 4 shows the average number of genes selected by each approach on the five real datasets. We can see that the genes selected by the penalized type methods (Lasso, SCAD and $L_{1/2}$) are significantly less than those of the non-penalized type methods (SPCR, PCR and PLS). The PLS and PCR methods selected all genes. Their performances are achieved at the expense of employing a quite large number of genes in the model, and are not suitable for knowledge discovery tasks. Among of the three penalized methods, the number of genes selected by the $L_{1/2}$ penalized method was the smallest, the SCAD was the second, and the Lasso was the largest.

Table 6
The CI results obtained by the six methods on the 5 real datasets. In bold – the best performance.

Datasets	SPCR	PCR	PLS	Lasso	SCAD	$L_{1/2}$
DLBCL(2002)	**0.5942**	0.5821	0.5506	0.5511	0.5517	0.5486
DLBCL(2003)	0.6001	0.6008	0.5898	0.6097	0.5841	**0.6121**
DBC	0.7004	0.7002	0.6994	0.7145	0.7209	**0.7214**
Lung cancer	0.5565	0.6149	0.5691	0.6236	**0.6355**	0.6246
AML	0.5861	0.5798	0.5989	**0.6117**	0.6019	0.6008
Mean	0.6074	0.6155	0.6015	0.6221	0.6188	0.6215

With respect to prediction accuracy, Tables 5 and 6 summarize the results of IBS and CI obtained by the six methods, respectively. In terms of IBS, the SPCR method performed a bit better than the other methods on the DLBCL(2002) dataset, the PCR method got the bit better performance on the DLBCL(2003) dataset, the Lasso method performed best on the AML dataset, the SCAD method performed best on the Lung cancer and AML datasets, the $L_{1/2}$ method performed best on the Dutch breast cancer dataset respectively. However, the results of all the methods were not much different. On the other hand, from terms of CI in Table 6, different methods yielded slightly larger CI value than the others on different datasets. For example, the SPCR method performed best with an CI of 0.5942 on the DLBCL(2002) dataset, and the $L_{1/2}$ method performed best on the DLBCL(2003) and Dutch breast cancer datasets, respectively. The predictive performances of those methods were similar and no one approach dominates the others. Combined with the results reported in Tables 4–6, we concluded that the prediction accuracies are comparable between all methods, and the $L_{1/2}$ penalized method selected the smaller subset of the prognostic genes, especially compared with the non-penalized methods. This is an important consideration for clinical applications, where the goal is often to develop an accurate predicting test using as few genes as possible in order to control cost. This indicates that if an analyst wants an extremely small subset of prognostic genes, the $L_{1/2}$ penalized method could be the best approach to use among these discussed methods.

5. Conclusion

Through extending the $L_{1/2}$ regularization theory in the linear regression framework to the nonlinear censored survival case, we introduced a new version of penalized Cox model based on the $L_{1/2}$ regularization. We developed the coordinate descent algorithm for the $L_{1/2}$ penalized Cox model with the Newton–Raphson iterative method. The proposed algorithm is applicable to biological data with high dimensions and low sample sizes.

Simulation results indicate that the $L_{1/2}$ penalized Cox model is very competitive in analyzing high dimensional survival data, because it was able to reduce the size of the predictor even further at moderate costs for the prediction accuracy. The $L_{1/2}$ penalized Cox model will provide an efficient tool in building a prediction model for survival time based on high dimensional biological data.

Acknowledgments

This research was supported by Macau Science and Technology Develop Funds (Grant No. 017/2010/A2) of Macau SAR of China and the National Natural Science Foundations of China (Grant Nos. 2013CB329404, 11131006, 61075054, and 11171272).

References

[1] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. B 34 (1972) 187–220.
[2] D.R. Cox, Partial likelihood, Biometrika 62 (1975) 269–762.
[3] E. Bair, R. Tibshirani, Semi-supervised methods to predict patient survival from gene expression data, PLoS Biol. 2 (2004) 511–522.
[4] E. Bair, et al., Prediction by supervised principal components, J. Am. Stat. Assoc. 101 (2006) 119–137.
[5] M.R. Segal, Microarray gene expression data with linked survival phenotypes: diffuse large B-cell lymphoma revisited, Biostatistics 7 (2006) 268–285.
[6] H.C. Van Houwelingen, T. Bruinsma, A.A.M. Hart, L.J. van't Veer, L.F.A. Wessels, Cross-validated Cox regression on microarray gene expression data, Stat. Med. 25 (2006) 3201–3216.
[7] R. Tibshirani, The lasso method for variable selection in the Cox model, Stat. Med. 16 (1997) 385–395.
[8] J. Fan, R. Li, Variable selection for Cox's proportional hazards model and frailty model, Ann. Stat. 30 (2002) 74–99.
[9] H.H. Zhang, W. Lu, Adaptive Lasso for Cox's proportional hazards model, Biometrika 94 (3) (2007) 691–703.
[10] D.A. Engler, Y. Li, Survival analysis with high-dimensional covariates: an application in microarray studies, Stat. Appl. Genet. Mol. Biol. 8 (2009) 14.
[11] Z.B. Xu, H. Zhang, Y. Wang, X.Y. Chang, $L_{1/2}$ regularization, Sci. China F 40 (3) (2010) 1–11.
[12] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B 58 (1996) 267–288.
[13] Z.B. Xu, X.Y. Chang, F.M. Xu, H. Zhang, $L_{1/2}$ regularization: a thresholding representation theory and a fast solver, IEEE Trans. Neural Netw. Learn. Syst. 23 (7) (2012) 1013–1027.
[14] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, Ann. Appl. Stat. 1 (2007) 302–332 (MR2415737).
[15] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.
[16] J. Gui, H. Li, Penalized Cox regression analysis in the high- dimensional and low-sample size settings, with applications to microarray gene expression data, Bioinformatics 21 (2005) 3001–3008.
[17] E. Graf, et al., Assessment and comparison of prognostic classification schemes for survival data, Stat. Med. 18 (1999) 2529–2545.
[18] F.E. Harrel, Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer, New York, NY, 2001.
[19] R. Bender, T. Augustin, M. Blettner, Generating survival times to simulate Cox proportional hazards models, Stat. Med. 24 (2005) 1713–1723.
[20] I. Sohn, J. Kim, S.H. Jung, C. Park, Gradient lasso for Cox proportional hazards model, Bioinformatics 25 (14) (2009) 1775–1781.
[21] T. Hastie, et al., The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.
[22] S. Nygard, et al., Partial least squares Cox regression on genomic data handling additional covariates, in: Statistical Research Report 5/2006, Department of Mathematics, University of Oslo, 2006 http://www.math.uio.no/eprint/stat_report/2006/05-06.html
[23] A. Rosenwald, et al., The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma, N. Engl. J. Med. 346 (2002) 1937–1946.
[24] A. Rosenwald, et al., The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma, Cancer Cell 3 (2003) 185–197.
[25] D.G. Beer, et al., Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (2002) 816–824.
[26] L. Bullinger, et al., Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia, N. Engl. J. Med. 350 (2004) 1605–1616.