Soft Comput (2014) 18:143–152
DOI 10.1007/s00500-013-1042-6

METHODOLOGIES AND APPLICATION

A novel $L_{1/2}$ regularization shooting method for Cox's proportional hazards model

Xin-Ze Luan · Yong Liang · Cheng Liu · Kwong-Sak Leung · Tak-Ming Chan · Zong-Ben Xu · Hai Zhang

Published online: 21 April 2013

Abstract Nowadays, a series of methods are based on a L_1 penalty to solve the variable selection problem for a Cox's proportional hazards model. In 2010, Xu et al. have proposed a $L_{1/2}$ regularization and proved that the $L_{1/2}$ penalty is sparser than the L_1 penalty in linear regression models. In this paper, we propose a novel shooting method for the $L_{1/2}$ regularization and apply it on the Cox model for variable selection. The experimental results based on comprehensive simulation studies, real Primary Biliary Cirrhosis and diffuse large B cell lymphoma datasets show that the $L_{1/2}$ regularization shooting method performs competitively.

Keywords Variable selection · Cox model · Lasso · $L_{1/2}$ regularization shooting algorithm

1 Introduction

One of the most important objectives for survival analysis is to select a small number of key risk factors from many potential predictors. Commonly, the Cox proportional hazards model (COX 1972, 1975) is used to study the relationship between predictor variables and survival time. Suppose a dataset has a sample size of n and we want to study the survival time t_i on covariate x, we represent the samples for an individual using (t_1, δ_1, x_1),…,(t_n, δ_n, x_n) where the survival time t_i being complete if $\delta_i = 1$ and right censored if $\delta_i = 0$. As in regression, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a vector of p potential predictors. We define $f(x) = \beta^T x$ to be the linear risk score function.

By the Cox's proportional hazards model, the hazard function is given as:

$$h(t|\beta) = h_0(t)\exp(\beta^T x) \tag{1}$$

where the baseline hazard function $h_0(t)$ is unspecified and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the regression coefficient vector of p variables.

In practice, not all the p predictors may contribute to the prediction of survival outcomes, i.e. some β 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, many variable selection techniques for linear regression models have been extended to the context of survival models. They include best-subset selection, stepwise selection, asymptotic procedures based on score tests, Wald tests and other approximate Chi squared testing procedures, bootstrap procedures (Sauerbrei and Schumacher 1992) and Bayesian variable selection (Faraggi and Simon 1998; Ibrahim et al. 1999). However, the theoretical properties of these methods are generally unknown (Fan and Li 2002).

Recently a series of penalized partial likelihood methods, such as the Lasso (Tibshirani 1996, 1997), the smoothly clipped absolute deviation method (Fan and Li

Communicated by G. Acampora.

X.-Z. Luan · Y. Liang (✉) · C. Liu
Macau University of Science and Technology, Macao, People's Republic of China
e-mail: yliang@must.edu.mo

K.-S. Leung · T.-M. Chan
Chinese University of Hong Kong, Hong Kong, People's Republic of China

Z.-B. Xu · H. Zhang
Xi'an Jiaotong University, Xi'an, People's Republic of China

2001, 2002) and the adaptive Lasso method (Zhang and Lu 2007), have been proposed for the Cox's proportional hazards model. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously. These series of the Lasso methods were based on the L_1 penalty. However, the L_1 type penalizations may not yield sufficiently sparse variable selection in real applications.

In this paper, we develop a novel shooting algorithm based on the $L_{1/2}$ regularization, which was proposed by Xu et al. (2010). It is shown that the $L_{1/2}$ regularization has many promising properties, such as unbiasedness, sparsity and oracle properties. The solution of the $L_{1/2}$ regularization is sparser than that of the L_1 regularization, while 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_p $(0 < p < 1)$ regularizations for the problems desiring sparsity. We use the $L_{1/2}$ regularization shooting algorithm to obtain the solutions for the Cox model in the setting of very high-dimensional covariates such as the gene expression data obtained by microarrays.

The rest of the paper is organized as follows. Section 2 present the Cox model and briefly review the L_1 type estimations of the regression coefficients and present the $L_{1/2}$ regularization approach. Section 3 gives a new shooting algorithm for obtaining the $L_{1/2}$ estimates. In Sect. 4, we evaluate the $L_{1/2}$ regularization shooting algorithm by simulation studies and applications to real datasets, such as the diffuse large B cell lymphoma (DLBCL) survival times and gene expression data. Finally, we give a brief discussion of the methods and conclusions in Sect. 5.

2 Related work

2.1 Regularization approaches for Cox proportional hazards model

Based on the available sample data, the Cox's partial log–likelihood (Cox 1972) can be written as

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ x_i^T \beta - \log\left(\sum_{j \in R_i} \exp(x_j^T \beta) \right) \right\} \tag{2}$$

where R_i denotes the set of indices of the individuals at risk at time t_i.

To select important variables under the proportional hazards model (2), Tibshirani (1997), Fan and Li (2002) and Zhang and Lu (2007) proposed to minimize the penalized log partial likelihood function,

$$-\frac{1}{n} l(\beta) + \lambda \sum_{j=1}^{p} P(\beta_j) \tag{3}$$

where $l(\beta)$ is the loss function, $P(\beta)$ is the penalty function and λ is the tuning parameter for variable selection. The series of the Lasso methods cannot directly be applied on the nonlinear Cox model to obtain parameter estimates. Therefore, Tibshirani (1997) and Zhang and Lu (2007) proposed iterative procedures to transform the Cox's partial log–likelihood function (2) to a linear regression problem through an iteratively Newton-Raphon update. Here we follow the approach of Zhang and Lu (2007): define the gradient vector $\nabla l(\beta) = -\partial l(\beta)/\partial \beta$ and the Hessian matrix $\nabla^2 l(\beta) = -\partial l^2(\beta)/\partial \beta \partial \beta^T$; then apply the Cholesky decomposition to obtain $\hat{x}^T = \left\{ \nabla^2 l(\beta) \right\}^{1/2}$; generate the pseudo response vector $\hat{y} = (\hat{x}^T)^{-1} \{ \nabla^2 l(\beta) \beta - \nabla l(\beta) \}$. By second-order Taylor expansion, $l(\beta)$ can be approximated by the quadratic form:

$$l(\beta) \approx (\hat{y} - \hat{x}\beta)^T (\hat{y} - \hat{x}\beta) \tag{4}$$

In the high dimensional part, we use another method to linearize the Cox model. Tibshirani (1997) proposed an iterative procedure to solve Eq. (2). Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, \ldots, n$, denote the $n \times p$ gene expression matrix, $\eta = \beta^T x$, $\mu = -\frac{\partial l}{\partial \eta}$, $A = -\frac{\partial^2 l}{\partial \eta \partial \eta^T}$, and $z = \eta + A^- \mu$, where A^- is a generalized inverse of A. By the Taylor expansion of $l(\beta)$, the partial log-likelihood is approximated by

$$l(\beta) \approx (z - \eta)^T A (z - \eta) \tag{5}$$

Tibshirani (1997) suggested to replace A with a diagonal matrix D having the same diagonal elements as λ and solve the formula (5) iteratively using a quadratic programming. Since the quadratic programming cannot be applied directly to the cases with $p \gg n$, Gui and Li (2005) applied the Cholesky decomposition to obtain $T = A^{1/2}$ such that $T^T T = A$, $\hat{y} = Tz$ and $\hat{x} = Tx$.

Thus at each iterative step, we can directly apply the Lasso linear regression on the approximated quadratic form the formulas (4) and (5). Tibshirani (1997) proposed to estimate parameters by the Lasso with the quadratic programming techniques:

$$\hat{\beta} = \arg\min \left\{ (\hat{y} - \hat{x}\beta)^T (\hat{y} - \hat{x}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{6}$$

we used above two linear regression methods of the Cox model in the simulation experiments to evaluate the performance of the $L_{1/2}$ regularization shooting method performs both in the low and high dimensional problems. We focus on whether the key coefficients that are related to survival endpoint can be selected by the $L_{1/2}$ regularization shooting method.

The L_1 penalization shrinks small coefficients to zero and hence results in a sparse representation of the solution.

However, the estimation of large β may suffer from substantial bias if λ is chosen too big and may not be sufficiently sparse if λ is selected too small. Hence, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty, which avoids excessive penalties on the large coefficients and enjoys the oracle properties. Gui and Li (2005) applied the LARS-Cox procedure for the Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Zhang and Lu (2007) suggested the adaptive Lasso method with an adaptively L_1 penalty $P(\beta) = \sum_{j=1}^{p} |\beta_j| / \left|\beta_j^{ols}\right|$ to estimate the parameters of the Cox model. Here the weights $1/\left|\beta_j^{ols}\right|$ are obtained by the ordinary linear regression. Compare to the Lasso penalty, the adaptive Lasso penalty can introduce different penalties to the different coefficients in a convex form and can be efficiently solved by the standard Lasso algorithms.

The above mentioned a series of the Lasso methods were based on the L_1 penalty. However, for many practical applications, the solutions of the L_1 regularization are often less sparse than those of the L_p $(0 \le p < 1)$ regularizations. To find sparser solutions than the L_1 regularization is, on the other hand, imperative and required for many real variable selection applications. Also, the L_1 regularization is inefficient when the errors in data have heavy tail distributions (Tibshirani 1996).

2.2 $L_{1/2}$ regularization

In this part, we introduce a $L_{1/2}$ regularization scheme for variable selection. Sparsity and parsimony of a statistical model is always desired, as the parsimonious models provide simple and interpretable relations among scientific variables in addition to reduce forecasting errors. A variety of variable selection criteria have been proposed. The best subset selection, namely, the L_0 penalty, along with the traditional model selection criteria such as AIC (Akaike 1973) and BIC (Schwarz 1978), involve solving a NP hard optimization problem, so they are infeasible for the high dimensional data. Consequently, an innovative variable selection procedure is expected to cope with very high dimensionality, which has been one of the hot topics in the field of machine learning. The regularization methods are recently developed as feasible approaches to solve this problem. In general, the regularization framework takes the form of the loss function $l(\beta)$ and the penalty function $P(\beta)$. Many existing learning algorithms can be considered as a special form of this regularization framework. For example, when the penalty function $P(\beta) = |\beta|^0$, it is AIC or BIC, which is referred to as the L_0 penalty in this paper. When the penalty function $P(\beta) = |\beta|$, it is the Lasso, which is called the L_1 penalty. When the penalty function $P(\beta) = \beta^2$, it is the ridge regression, which is called the L_2 penalty. And when the penalty function $P(\beta) = |\beta|^\infty$, it is the L_∞ penalty.

The L_0 penalty is the earliest regularization method applied to variable selection and feature extraction. Constrained by the number of coefficients including nonzero, the L_0 penalty yields the sparsest solutions, but it has to solve a NP hard combinatory optimization problem. The L_1 penalty (Lasso) proposed by Tibshirani (1996) provides an iteration for variable selection and feature extraction, which just needs to solve a quadratic programming problem but is less sparse than the L_0 penalty. At the same time, Donoho and Huo (2001), Donoho and Elad (2003) and Chen and Donoho (2001) proposed Basis Pursuit when studying the signal sparsity recovery problem. They proved that under some conditions the solutions of the L_0 penalty are equivalent to those of the L_1 penalty for the sparsity problem, so the NP hard optimization problem can be avoided by applying the L_1 penalty. Based on their works, the L_1 penalty, or more generally, the L_1 type penalties, including SCAD (Fan and Heng 2004), the adaptive Lasso (Zou 2006; Zhang and Lu 2007), Elastic net (Zou and Hastie 2005), Stagewise Lasso (Zhao and Yu 2007), and Dantzig selector (Candes and Tao 2007), have become the dominant tools for data analysis since then.

In recent years, Xu et al. (2010) proposed the $L_{1/2}$ regularization:

$$\beta_{1/2} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \hat{x}_i \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^{1/2} \right\} \tag{7}$$

where λ is the tuning parameter. Different from the L_1 penalty, the $L_{1/2}$ regularization is nonconvex. Xu et al. (2010) have proved the properties of sparsity, unbiasedness and oracle properties of the $L_{1/2}$ penalty. Their experiments show that the solutions yielded from the $L_{1/2}$ penalty are sparser and can predicate better than those from the L_1 penalty. On the other hand, solving the $L_{1/2}$ penalty is much simpler than solving the L_0 penalty. All these properties support the usefulness of the $L_{1/2}$ penalty and the $L_{1/2}$ penalty can be potentially more powerful than the L_0 and L_1 penalties in real applications.

3 A novel shooting algorithm for the $L_{1/2}$ regularization

In literatures, Fu (1998) investigated the Lasso Shooting algorithms based on the L_1 penalty. According to Xu et al.'s (2010) research achievement, the $L_{1/2}$ regularization can be transformed into that of a series of convex weighted

Lasso. Here, we propose a new shooting algorithm for the $L_{1/2}$ regularization, and its procedure is as follow:

Step 1: Initialize $t = 1$ and $\beta^0 = (\underbrace{0, 0, \cdots 0}_{p})$

Step 2: Compute∇l,$\nabla^2 l$, $\hat{x}$ and $\hat{y}$ based on the current value β^{t-1}.

Step3: At step t, for each $j = 1,\ldots p$, and set

$$\beta_j^t = \begin{cases} \frac{\lambda\left|\beta_j^{t-1}\right|^{-\frac{1}{2}}-2F_0}{4(\hat{x}_j)^T\hat{x}_j} & \text{if } F_0 > \frac{1}{2}\lambda\left|\beta_j^{t-1}\right|^{-\frac{1}{2}} \\ \frac{-\lambda\left|\beta_j^{t-1}\right|^{-\frac{1}{2}}-2F_0}{4(\hat{x}_j)^T\hat{x}_j} & \text{if } F_0 < \frac{1}{2}\lambda\left|\beta_j^{t-1}\right|^{-\frac{1}{2}} \\ 0 \quad \text{if } |F_0| \le \frac{1}{2}\lambda\left|\beta_j^{t-1}\right|^{-\frac{1}{2}} \end{cases}$$

where $\hat{x}_j$ is the jth column vector of $\hat{x}$, and λ is the tuning parameter. Define $RSS = (\hat{y} - \hat{x}\beta^{t-1})^T(\hat{y} - \hat{x}\beta^{t-1})$, $F_j(\beta^{t-1}, \hat{x}, \hat{y}) = \frac{\partial RSS}{\partial \beta_j^{t-1}}$, $F_0 = F_j(\beta^{t-1,-j}, \hat{x}, \hat{y})$, use $\beta^{t-1,-j}$ to denote $(\beta_1^{t-1}, \beta_2^{t-1}, \ldots, \beta_{j-1}^{t-1}, 0, \beta_{j+1}^{t-1}, \ldots, \beta_p^{t-1})^T, j = 1,\ldots, p$, and form a new estimator$\beta^t = (\beta_1^t, \ldots, \beta_p^t)^T$after updating all $\beta_j^t (j = 1,\ldots, p)$.

Step 4: Let $t = t + 1$. Go back to Step 2 until the following convergence criterion is satisfied:

$$\sum_{i=1}^{p} \left|\beta_i^t - \beta_i^{t-1}\right| < 10^{-5}.$$

To determine the value of the tuning parameter λ, we use the maximization of the cross validated partial likelihood (CVPL) (Verwij and Van Houwelingen 1993; Huang and Harrington 2002), which is defined as

$$CVPL(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\left[l(\hat{f}^{(-i)}(\lambda)) - l^{(-i)}(\hat{f}^{(-i)}(\lambda))\right] \tag{8}$$

where $\hat{f}^{(-i)}(\lambda)$ is the estimation of the score function based on the $L_{1/2}$ procedure with the tuning parameter λ from the data without the ith subject. The terms $l(f)$ and $l^{(-i)}(f)$ are the log partial likelihoods with all the subjects and without the ith subject, respectively. The optimal value of λ is chosen to maximize the sum of the contributions of each subject to the log partial likelihood. CVPL is the special case of a more general cross-validated likelihood approach for model selection (Smyth 2001; Van der Laan et al. 2003) and has been demonstrated to perform well in prediction in the context of the penalized Cox regression (Huang and Harrington 2002).

In Xu et al.'s (2010) paper, the convergence of $L_{1/2}$ penalty algorithms has been proved. They show that the $L_{1/2}$ regularization algorithms will always approach to the set of global or local minima of the problems.

4 Numerical studies

4.1 The low-dimensional simulation for the Cox model

In this part, we compare the performance of the Lasso, the adaptive Lasso, and the $L_{1/2}$ regularization shooting algorithm under the Cox's proportional hazards model. The cross validated partial likelihood (CVPL) method is used to estimate the tuning parameter λ in these three algorithms. To report the estimation bias for the true predictor variables of the three methods, we follow Tibshirani (1997) and summarize the average mean squared errors $(\hat{\beta} - \beta)^T V(\hat{\beta} - \beta)$ over many runs. Here V is the population covariance matrix of the covariates.

To measure the prediction accuracy, Graf et al. (1999) have proposed to use the time-dependent Brier (1950) score (BS), which is the time-dependent mean-squared error between the observed survival status and the predicted survival probability. The BS depends on time t. Thus it makes sense to use the integrated Brier score (IBS) as a score to assess the goodness of the predicted survival functions of all observations between time 0 and an arbitrary upper limit t^* (Graf et al. 1999).

In our simulation studies, we selected the Gompertz model which is frequently used in human mortality model and has the property of proportion of hazards, here we following the method of Qian et al. (2010) to generate the datasets of the Cox model. Detailed steps of generating survival data with the censoring rate are described as follows:

Step 1: The survival time T_i ($i = 1,\ldots, n$, n indicates sample size) is constructed from a uniformly distributed variable U by $T_i = \frac{1}{\alpha}\log\left(1 - \frac{\alpha\times\log(U)}{\gamma\exp(x_i\beta)}\right)$, where γ is scale parameter, α is shape parameter, β is the ground-true regression coefficient and the covariates x_i is p dimensional and normally distributed vector with different parameter settings.

Step 2: 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 3: Define $t_i = \min(T_i, T_i')$ and $\delta_i = I(T_i \le T_i')$. Therefore, we can generate the observed data consist of (t_i, δ_i, x_i) for Cox proportional hazards model.

For our experiment, we generated simulation datasets in two setting.

Model 1: $\beta = (-0.7, 0, -0.7, 0, 0, 0, 0, -0.7, 0)$, where important variables have large effects;

Model 2: $\beta = (-0.3, 0, -0.2, 0, 0, 0, 0, -0.1, 0)$, where important variables have small effects.

We considered two censoring rates, 25 and 40 % and three samples sizes $n = 150$, 250, 350 respectively.

The average numbers of zero coefficients obtained by the three methods are reported in the Table 1. From Table 1, the $L_{1/2}$ regularization shooting method performs best in terms of both variable selection and prediction accuracy. For example, in the Corr columns for Model 1, when $n = 150$ and the censoring is 25 %, where the true model has 6 zero coefficients, the average numbers of the correct zero coefficients from the Lasso is 4.78, from the adaptive Lasso is 5.43 and from the $L_{1/2}$ regularization shooting method is 5.79. This means that the $L_{1/2}$ regularization shooting method shrinks unimportant covariates most accurately. Moreover, the mean squared errors (MSE) of the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting method are 0.1375, 0.0629 and 0.0582 (best). The IBS's values of these three methods are 0.1131, 0.1140 and 0.1124 respectively. It means that the $L_{1/2}$ regularization shooting method performs slight better than the other two methods for the prediction accuracy. As n increases to 250 or 350, the performance of the $L_{1/2}$ regularization shooting method is still consistently better than those of other two methods. In the Incorr columns, the idealized average number is 0 if the method can correctly identify all relevant variables at each run, whereas, its maximal value is 3 if the method incorrectly identifies all the nonzero coefficients of the relevant variables to zero in all runs. From the Incorr columns, we can also find that all the three algorithms never evaluated the nonzero coefficients to zero. Similar results are observed for the 40 % censoring case.

In Model 2, the important variables have small effects and its coefficients are of different magnitudes. The second part of Table 1 shows that the $L_{1/2}$ regularization shooting method is best in terms of shrinking non-important covariates under the different parameter settings. For example, when $n = 150$ and the censoring is 25 %, in the Corr columns, the average numbers of the correct zero coefficients from the Lasso is 3.79, from the adaptive Lasso is 5.26 and from the $L_{1/2}$ regularization shooting method is 5.57. The correct number of zeros is 6. In regard to prediction accuracy, the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting method give similar IBS values and performance under the different parameter settings.

In the Incorr columns, each method performs well in Model 1. However, in Model 2, when n = 150 and censoring is 25 %, for the three relevant variables, the average number of the incorrect zeros from the Lasso is 0.35, from the adaptive Lasso is 0.61 and from the $L_{1/2}$ regularization

Table 1 The simulation results based on the Models 1 and 2 by the three methods over 100 replications. The columns include the average number of the correct zeros (Corr), the average number of the incorrect zeros (Incorr), the mean squared error (MSE) and the integrated Brier score (IBS)

n	25 % censoring					40 % censoring			
	Method	Corr (6)	Incorr (0)	MSE	IBS	Corr (6)	Incorr (0)	MSE	IBS
Model 1:$\beta = [-0.7, 0, -0.7, 0, 0, 0, 0, -0.7, 0]$									
150	Lasso	4.78	0.0	0.1375	0.1131	4.46	0.0	0.0965	0.1137
	Adaptive	5.43	0.0	0.0629	0.1140	5.50	0.0	0.0362	0.1147
	$L_{1/2}$	5.79	0.0	0.0582	0.1124	5.69	0.0	0.0307	0.1132
250	Lasso	4.63	0.0	0.1955	0.1123	4.41	0.0	0.0939	0.1108
	Adaptive	5.62	0.0	0.0440	0.1132	5.58	0.0	0.0234	0.1118
	$L_{1/2}$	5.80	0.0	0.0381	0.1114	5.79	0.0	0.0216	0.1101
350	Lasso	4.81	0.0	0.1857	0.1108	4.51	0.0	0.1007	0.1094
	Adaptive	5.63	0.0	0.0323	0.1116	5.69	0.0	0.0190	0.1103
	$L_{1/2}$	5.90	0.0	0.0308	0.1098	5.90	0.0	0.0183	0.1086
Model 2 :$\beta = [-0.3, 0, -0.2, 0, 0, 0, 0, -0.1, 0]$									
150	Lasso	3.79	0.35	0.0400	0.1271	3.29	0.16	0.0336	0.1276
	Adaptive	5.26	0.61	0.0448	0.1272	5.37	0.45	0.0368	0.1278
	$L_{1/2}$	5.57	0.83	0.0410	0.1269	5.17	0.42	0.0328	0.1275
250	Lasso	4.22	0.22	0.0375	0.1230	3.36	0.02	0.0253	0.1289
	Adaptive	5.44	0.43	0.0347	0.1231	5.42	0.19	0.0242	0.1291
	$L_{1/2}$	5.62	0.42	0.0312	0.1228	5.43	0.15	0.0213	0.1288
350	Lasso	4.60	0.07	0.0340	0.1231	3.39	0.0	0.0245	0.1218
	Adaptive	5.41	0.33	0.0239	0.1234	5.41	0.37	0.0213	0.1220
	$L_{1/2}$	5.79	0.28	0.0229	0.1229	5.41	0.25	0.0182	0.1217

Lasso the Lasso method, *Adaptive* the adaptive Lasso method, $L_{1/2}$ the $L_{1/2}$ regularization shooting method

shooting method is 0.83. The $L_{1/2}$ regularization shooting method performs worse than other two methods. As n increases to 250 or 350, the Lasso method still performs best, but the $L_{1/2}$ regularization shooting method is better than the adaptive Lasso method. Similar results are observed for the 40 % censoring case. This means that the adaptive Lasso and the $L_{1/2}$ regularization shooting methods shrink the small effect covariates to zero more easily than the Lasso.

4.2 Experiments on the real PBC (primary biliary cirrhosis) dataset

We follow Tibshirani (1997) and Zhang and Lu (2007) applications to PBC dataset. The data is from the Mayo Clinic trial in primary biliary cirrhosis of liver conducted between 1974 and 1984, and provided in Dickson et al. (1989), a more detailed account can be found in Therneau and Grambsch (2000). In this study, 312 out of 424 patients who agreed to participate in the randomized trial are eligible for the analysis. For each patient, clinical, biochemical, serological and histological parameters are collected. In all, 125 patients died before the end of follow-up. We study the dependence of the survival time on seventeen covariates: continuous variables are age (in years), alb (albumin in g/dl), alk (alkaline phosphatase in units/litre), bil (serum bilirubin in mg/dl), chol (serum cholesterol in mg/dl), cop (urine copper in g/day), plat (platelets per cubic ml/1000), prot (prothrombin time in seconds), sgot (liver enzyme in units/ml), trig (triglycerides in mg/dl); categorical variables are asc (0 denotes absence of ascites and 1 denotes presence of ascites), oed (0 denotes no oedema, 0.5 denotes untreated or successfully treated oedema and 1 denotes unsuccessfully treated oedema), hep (0 denotes absence of hepatomegaly and 1 denotes presence of hepatomegaly), sex (0 denotes male and 1 denotes female), spid (0 denotes absence of spiders and 1 denotes presence of spiders), stage (histological stage of disease, graded 1, 2, 3 or 4) and trt (1 for control and 2 for treatment). We focus only on the 276 observations without missing values. All seventeen variables are included in the model. Table 2 gives the estimated coefficients by the three methods. As reported in Tibshirani (1997), the stepwise selection chooses eight variables, namely age, oed, bil, alb, cop, sgot, prot and stage, which are similar to the results of the adaptive Lasso and the $L_{1/2}$ regularization shooting methods (eight variables) and better than that of the Lasso (nine variables). In regard to prediction accuracy, the IBS's values of these three methods are 0.1874, 0.1877 and 0.1875 respectively in Table 3, which means that they have similar prediction performances.

Table 2 Estimate coefficients by the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting method for Primary biliary cirrhosis data

Covariate	Lasso	Adaptive	$L_{1/2}$
trt	0.0	0.0	0.0
age	0.3059 (0.0021)	0.4212 (0.0024)	0.4895 (0.00057)
sex	0.0	0.0	0.0
asc	0.0401 (0.0002)	0.0	0.0
hep	0.0	0.0	0.0
spid	0.0	0.0	0.0
oed	0.3021 (0.0006)	0.2994 (0.0008)	0.3348 (0.0002)
bil	0.6411 (0.0000)	0.7071 (0.0016)	0.6624 (0.0001)
chol;	0.0	0.0	0.0
alb	−0.3843 (0.0009)	−0.6396 (0.0021)	−0.4668 (0.0001)
cop	0.4072 (0.0001)	0.4471 (0.0012)	0.4185 (0.0000)
alk	0.0	0.0	0.0
sgot	0.1481 (0.0024)	0.1820 (0.0059)	0.3333 (0.0008)
trig	0.0	0.0	0.0
palt	0.0	0.0	0.0
prot	0.2325 (0.0013)	0.2839 (0.0019)	0.3332 (0.0005)
stage	0.4043 (0.0016)	0.5083 (0.0016)	0.5709 (0.0004)

The standard deviations are given in parentheses. (*Lasso* the Lasso method, *Adaptive* the adaptive Lasso method, $L_{1/2}$ the $L_{1/2}$ regularization shooting method)

Table 3 The integrated Brier score (IBS) obtained by the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting method for PBC data

	Lasso	Adaptive	$L_{1/2}$
IBS	0.1874	0.1877	0.1875

Lasso the Lasso method, *Adaptive* the adaptive Lasso method, $L_{1/2}$ the $L_{1/2}$ regularization shooting method

4.3 The high-dimensional experiment for the Cox model

In our high-dimensional simulation for the Cox model, we assume that 1,000 genes are considered from the above Gompertz model method which was introduced in the part 1 of this section. Among of them, six genes are related to time to cancer recurrence. In our experiment, we generated simulation datasets in the setting:

$$\beta = \left(\underbrace{\overbrace{-0.7, -0.6, -0.5, -0.4, 0, 0, 0, 0, 0, 0, 0.4, 0, 0, 0.7}^{14}, \overbrace{0, \ldots, 0}^{986}}_{1{,}000} \right)$$

We considered the cases with 25 and 40 % of censoring and used four samples, $n = 200$, 250, 300, 350. The simulation results obtained by the three methods reported in Table 4. Since this simulation dataset has six relevant

Table 4 The simulation results based on the high dimensional simulated dataset by the three methods over 50 replications. The columns include the average number of the selected variable (Var), the average number of the correct zeros (Corr), the average number of the incorrect zeros (Incorr), and the integrated Brier score (IBS)

n	25 % censoring					40 % censoring			
	Method	Var	Corr(994)	Incorr(0)	IBS	Var	Corr(994)	Incorr(0)	IBS
200	Lasso	78.62	921.22	0.16	0.1348	74.30	925.54	0.16	0.1405
	Adaptive	31.54	968.22	0.24	0.1326	23.20	976.60	0.20	0.1382
	$L_{1/2}$	19.61	980.07	0.32	0.1311	16.10	983.62	0.28	0.1371
250	Lasso	99.45	900.55	0.00	0.1346	98.66	901.34	0.00	0.1372
	Adaptive	41.03	958.89	0.08	0.1328	32.60	967.34	0.06	0.1354
	$L_{1/2}$	24.47	975.37	0.16	0.1312	19.94	979.98	0.08	0.1342
300	Lasso	124.66	875.34	0.00	0.1328	105.60	894.40	0.00	0.1346
	Adaptive	59.76	940.24	0.00	0.1313	41.36	958.64	0.00	0.1331
	$L_{1/2}$	29.20	970.78	0.02	0.1298	23.84	976.12	0.04	0.1318
350	Lasso	156.92	843.08	0.00	0.1313	126.40	873.60	0.00	0.1343
	Adaptive	78.36	921.64	0.00	0.1300	52.44	947.56	0.00	0.1330
	$L_{1/2}$	33.44	966.56	0.00	0.1286	27.36	972.64	0.00	0.1317

Lasso the Lasso method, *Adaptive* the adaptive Lasso method, $L_{1/2}$ the $L_{1/2}$ regularization shooting method

features (six nonzero coefficients) in the 1,000 ones, the idealized average numbers of variables selected (the Var column) and correct zeros (the Corr column) by each method are 6 and 994 respectively. From the Var and Corr columns of Table 4, the results obtained by the $L_{1/2}$ regularization method are obviously better than those of other methods for different sample sizes and censoring settings. For example, when $n = 200$ and the censoring is 25 %, the average numbers (Var) from the Lasso, the adaptive Lasso, and the $L_{1/2}$ regularization methods are 78.62, 31.54 and 19.61 (best). The correct zeros' numbers (Corr) of the three methods are 921.22, 968.22 and 980.07 (best) respectively. The results obtained by the $L_{1/2}$ method are obviously close to the idealized values in the Var and Corr columns. Moreover, in the IBS column, the IBS's value of the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting method are 0.1348, 0.1326 and 0.1311. This means that the $L_{1/2}$ regularization shooting method performs slight better than the other two methods for the prediction accuracy. Similar results are observed for the 40 % censoring case.

As shown in the Incorr columns of Table 4, the idealized average number is 0 if the method can correctly identify all relevant variables at each run, whereas, its maximal value is 6 if the method incorrectly identifies all the nonzero coefficients to zero in all runs. When the sample size is relative small (n = 200 and censoring rate = 25 %), the average number of the incorrect zeros from the Lasso is 0.16, from the adaptive Lasso is 0.24 and from the $L_{1/2}$ regularization shooting method is 0.32. The $L_{1/2}$ regularization shooting method performs worse than other two methods. When n increases to 350, all the three algorithms never evaluated the nonzero coefficients to zero. This means that the $L_{1/2}$ regularization shooting method shrinks the small effect covariates to zero more easily than the Lasso and the adaptive Lasso when the sample size is relative small. Similar results are observed for the 40 % censoring case.

4.4 Experiments on the high-dimensional and real DLBCL (diffuse large B cell lymphoma) dataset

To further demonstrate the utility of the $L_{1/2}$ regularization shooting procedure in relating microarray gene expression data to censored survival phenotypes, we re-analyzed a published dataset of DLBCL by Rosenwald et al. (2002). This dataset contains a total of 240 patients with DLBCL, including 138 patient deaths during the follow-ups with a median death time of 2.8 years. Rosenwald et al. (2002) divided the 240 patients into a training set of 160 patients and a test set of 80 patients and built a multivariate Cox model. The variables in the Cox model included the average gene expression levels of smaller sets of genes in four different gene expression signatures together with the gene expression level of BMP6. It should be noted that in order to select the gene expression signatures, they performed a hierarchical clustering analysis for genes across all the samples (including both training and test samples). In order to compare our results with those in Rosenwald et al. (2002), we used the same setting of training and test datasets in our analysis.

We applied the $L_{1/2}$ regularization shooting method to first build a predictive model using the training data of 160

Table 5 GenBank ID and descriptions of the top 9 genes selected by the $L_{1/2}$ regularization method based on the 160 patients in the training dataset

GenBankID	Signature	Description
NM_005191		
AA714513	MHC	Major histocompatibility complex, class II, DR beta 5
AA767112	MHC	Major histocompatibility complex, class II, DP beta 1
X82240		
AA805575	Germ	Thyroxine-binding globulin precursor
AA505045	Germ	Homo sapiens, clone MGC:3963 IMAGE:3621362, mRNA, complete cds
AA598653	Lymph	Osteoblast specific factor 2 (fasciclin I-like)
AA598653	Lymph	Secreted protein, acidic, cysteine-rich (osteonectin)
LC_24433	Lymph	

As indicated are the gene expression signature groups that these genes belong to; *Germ* germinal-center B cell signature, *MHC* MHC class II signature, *Lymph* lymph-node signature. Genes NM_005191 and X82240 do not belong to these signature groups

patients and all the 7,399 genes as features (predictors). Table 5 shows the GenBank ID and a brief description of top nine genes selected by our proposed $L_{1/2}$ regularization method. It is interesting to note that seven of these genes belong to the gene expression signature groups defined in Rosenwald et al. (2002). These three signature groups include Germinal-center B cell signature, MHC, and Lymph-node signature. On the other hand, two genes selected by the $L_{1/2}$ method are not in the proliferation signature group defined by Rosenwald et al. (2002).

Based on the estimated model with these genes, we estimated the risk scores using the method proposed by Gui and Li (2005). To further examine whether clinically relevant groups can be identified by the model, we used zero as a cutoff point of the risk scores and divided the test patients into two groups based on whether they have positive or negative risk scores ($f(x) = \beta^T x$).

As a comparison, the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization methods are validated on the test dataset of 80 patients defined in Rosenwald et al. (2002), and their corresponding Kaplan–Meier curves are shown in Fig. 1. In Fig. 1, the horizontal coordinate is the predictive survival time (years) and the vertical coordinate is the predictive survival probabilities. The p values (lower the better to indicate statistical significance) of the Lasso for the test dataset is 0.0007, which are significantly larger than those of the adaptive Lasso and the $L_{1/2}$ regularization methods. This means that lasso method performs the worst for the survival prediction compared with other two methods.

On the other hand, in order to assess how well the model predicts the outcome, we also use the idea of receiver-operator characteristics (ROC) curves for the test dataset including censored observations and the area under the curve (AUC) as our criteria. These methods were developed by Heagerty et al. (2000) in the context of the medical diagnosis. Figure 2 shows the specific time ROC curves corresponding to the three methods at the beginning of the following-ups. We can see that the AUC's values from the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting methods are 0.702, 0.726 and 0.716 respectively. Note that a larger AUC at time t indicates better predictability of time to event at time t as measured by sensitivity and specificity evaluated at time t. It means that for the DLBCL datasets, the lasso method performs the worst than the other two methods.

In Table 6, the IBS's value of the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting method are 0.2188, 0.2040 and 0.2049. We can see that the adaptive Lasso and the $L_{1/2}$ regularization shooting methods perform slight better than Lasso for the prediction accuracy.

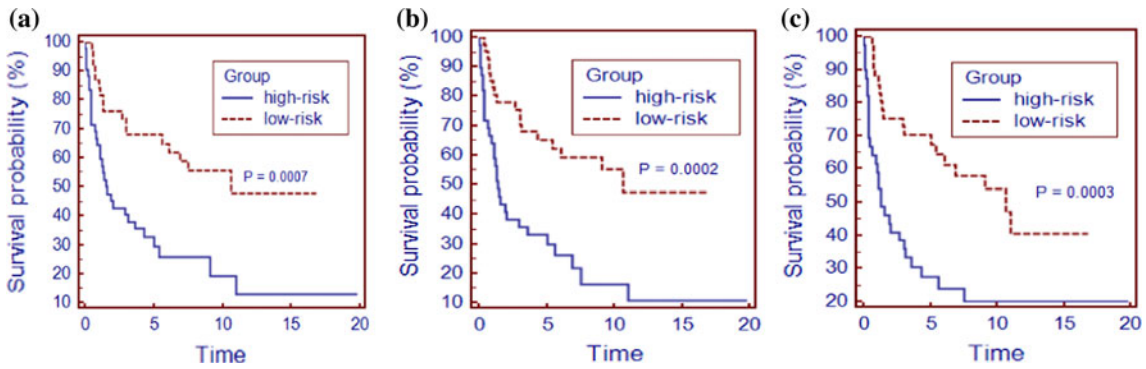
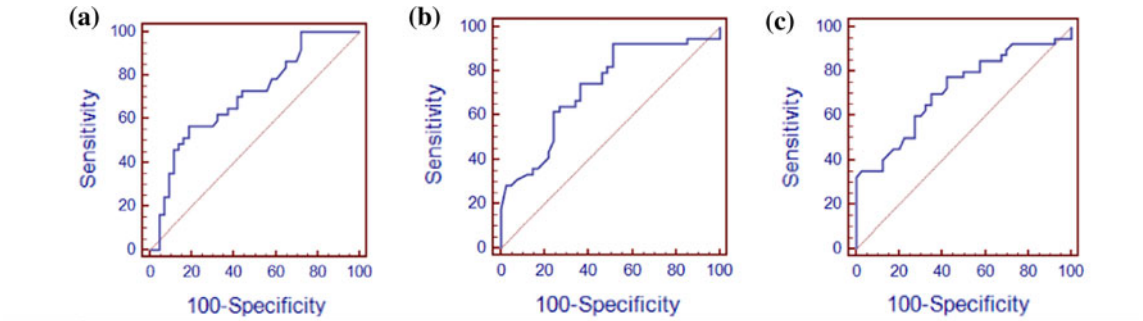


Fig. 1 The Kaplan–Meier *curves* for the high and low risk groups defined by the estimated scores for the 80 patients in the test dataset. The scores are estimated based on the models estimated by the Lasso method (plot **a**), the adaptive Lasso method (plot **b**) and the $L_{1/2}$ regularization shooting method (plot **c**). The maximal follow-up time is 20 years



Area under the ROC curve (AUC)	0.702	0.726	0.716
Standard Error[a]	0.0595	0.0574	0.0579
95% Confidence interval[b]	0.589 to 0.799	0.615 to 0.820	0.604 to 0.811

Fig. 2 The ROC *curves* at the specific time ($t = 0$) after diagnosis based on the estimated scores for the 80 patients in the test dataset. The scores are estimated based on the models estimated by the Lasso method (plot **a**), the adaptive Lasso method (plot **b**) and the $L_{1/2}$ regularization shooting method (plot **c**)

Table 6 The integrated Brier score (IBS) obtained by the Lasso, the adaptive Lasso and the $L_{1/2}$ regularization shooting method for DLBCL dataset

	Lasso	Adaptive	$L_{1/2}$
IBS	0.2188	0.2040	0.2049

Lasso the Lasso method, *Adaptive* the adaptive Lasso method, $L_{1/2}$ the $L_{1/2}$ regularization shooting method

5 Conclusion

In this paper, we have presented a novel $L_{1/2}$ regularization shooting method, which is used for variable selection in the Cox's proportional hazards model. Its performance is validated by both simulation and real case studies. In the experiments, we use two real datasets. One of the datasets is low-dimensional and high-sample size settings ($n > p$), from the result we can see that the $L_{1/2}$ regularization shooting algorithm performs better than the Lasso and the adaptive Lasso methods for variable selection and prediction accuracy. The other dataset is the high-dimensional and low-sample size settings, with applications to microarray gene expression data ($n << p$, DLBCL). Results indicate that our proposed $L_{1/2}$ regularization shooting algorithm is very competitive in analyzing high dimensional survival data in terms of sparsity of the final prediction model and predictability. The proposed $L_{1/2}$ regularization procedure is very promising and useful in building a parsimonious predictive model used for classifying future patients into clinically relevant high-risk and low-risk groups based on the gene expression profile and survival times of previous patients. The procedure can also be applied to select important genes which are related to patient's survival outcome.

Acknowledgments This work was supported by the Macau Science and Technology Develop Funds (Grant No. 017/2010/A2) of Macau SAR of China and the National Natural Science Foundation of China (Grant No. 11171272).

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Caki F (eds) Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp 267–281

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3

Candes E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35:2313–2351

Chen S, Donoho DL, Saunders M (2001) Atomic decomposition by basis pursuit. SIAM Rev 43:129–159

Cox DR (1972) Regression models and life-tables. J R Statist Soc B 34:187–220

Cox DR (1975) Partial likelihood. Biometrika 62:269–276

Dickson E, Grambsch P, Fleming T, Fisher L, Langworthy A (1989) Prognosis in primary biliary cirrhosis: model for decision making. Hepatology 10:1–7

Donoho DL, Elad E (2003) Maximal sparsity representation via l1 minimization. Proc Natl Acal Sci 100:2197–2202

Donoho DL, Huo X (2001) Uncertainty principles and ideal atomic decomposition. IEEE Trans Inf Theory 47:2845–2862

Fan J, Heng P (2004) Nonconcave penalty likelihood with a diverging number of parameters. Ann Stat 32(2):928–961

Fan J, LI R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Fan J, LI R (2002) Variable selection for Cox's proportional hazards model and frailty model. Ann Stat 30:74–99

Faraggi D, Simon R (1998) Bayesian variable selection method for censored survival data. Biometrics 54:1475–1485

Fu W (1998) Penalized regression: the bridge versus the lasso. J Comp Graph Stat 7:397–416

Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18:2529–2545

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 21:3001–3008

Heagerty PJ et al (2000) Time dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56:337–344

Huang J, Harrington D (2002) Penalized partial likelihood regression for right censored data with bootstrap selection of the penalty parameter. Biometrics 58:781–791

Ibrahim JG, Chen M-H, Maceachern SN (1999) Bayesian variable selection for proportional hazards models. Can J Stat 27:701–717

Qian J, Li B, Chen PY (2010) Generating survival data in the simulation studies of cox model. Inf Comput (ICIC) 4:93–96

Rosenwald A et al (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. New Engl J Med 346:1937–1946

Sauerbrei W, Schumacher M (1992) A bootstrap resampling procedure for model building: application to the cox regression model. Stat Med 11:2093–2109

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Smyth P (2001) Model selection of probabilistic clustering using cross-validated likelihood. Stat and Comput 10:63–72

Therneau TM, Grambsch PM (2000) Modeling survival data: Extending the Cox model. Springer-Verlag Inc., NewYork

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288

Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16:385–395

Van Der Laan MJ, Dudoit S, Keles S (2003) Asymptotic Optimality of likelihood based Cross Validation, Technical Report, Division of Biostatistics, University of California, Berkeley

Verwij PJM, Van Houwelingen JC (1993) Cross validation in survival analysis. Stat Med 12:2305–2314

Xu ZB, Zhang H, Wang Y, Chang XY (2010) $L_{1/2}$ regularization. Sci China, ser F 40(3):1–11

Zhang HH, Lu W (2007) Adaptive Lasso for Cox's proportional hazards model. Biometrika 94:691–703

Zhao P, Yu B (2007) Stagewise Lasso. J Mach Learn Res 8:2701–2726

Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B 67:301–320