A Gradient Lasso Algorithm for Cox Proportional Hazard Model

Insuk Sohn a, Jinseog Kim b, Changyi Park c, Sin-Ho Jung a

a Department of Biostatistics & Bioinformatics, Duke University, NC 27705, USA; b Department of Statistics and Information Science, Dongguk University, Gyeongju, 780-714; c Department of Statistics, University of Seoul, Seoul 130-743, Korea

ABSTRACT

Motivation: There has been an increasing interest in relating gene expression profiles to survival phenotypes such as time to cancer recurrence or time to death. Due to high dimensionality of gene expression data, there is a serious problem of collinearity in applying Cox regression directly. To avoid the collinearity problem, several methods based on penalized Cox regression have been proposed. However proposed methods suffer from severe computational problems in their optimization. We propose to implement penalized Cox regression with lasso penalty via gradient Lasso algorithm which yields faster convergence to the global optima than other algorithms. Our Cox regression model can deal with gene expressions as well as clinical covariates at the same time. By applying LASSO penalty only to gene expressions, our final model consists of relevant genes and clinical covariates. We propose a generalized cross validation type statistic for the choice of the tuning parameter for the lasso penalty.

Results:
Availability: park463@uos.ac.kr

1 INTRODUCTION

The DNA microarray is a new tool in biotechnology allowing the simultaneous monitoring of thousands of gene expression in cells (?) ... It has important applications in pharmaceutical and clinical research, including tumor classification, molecular pathway modeling, and functional genomics. The identification of genes that correlate with survival may provide new information on pathogenesis and etiology, and may further aid in the search for new targets for drug design.

There has been an increasing interest in relating gene expression profiles to survival phenotypes such as time to cancer recurrence or time to death. Recent work includes semi-supervised methods (?), supervised principal components (?), Cox regression based on threshold gradient descent (?), LARS-Cox (?), residual finesse (?), cancer survival prediction using automatic relevance determination (?), and sparse kernel methods.

In particular, we are interested in the Cox regression (?) because it is the most popular method for survival data with censoring. However, due to high dimensionality of gene expression data, there is a serious problem of collinearity in applying Cox regression directly. To avoid the collinearity problem, several methods based on penalized Cox regression have been proposed. ? adopted the smoothly-clipped-absolute-deviation penalty in the Cox model and ? proposed a path-based variable selection method to construct an adaptive sparse shrinkage path of the coefficients. To speed up the computation, ? adopted a quadratic approximation of the partial likelihood. ? suggested an adaptive lasso method with adaptively weighted penalty on coefficients. LARS-Cox procedure by ? is known to be the most effective method. However, LARS-Cox procedure suffers from severe computational problems in its optimization. ? proposed a method, called the residual finesse, to speed-up the computation by replacing the survival times by the deviance residuals.

In this paper, we propose to implement penalized Cox regression with lasso penalty via gradient lasso algorithm by ?, which yields faster convergence to the global optima than others. In addition, we propose a generalized cross validation (GCV) type statistic for the choice of the tuning parameter for the lasso penalty. The gradient lasso algorithm together with GCV type statistic is expected to speed-up the computation of the penalized Cox regression without loss of predictive performance.

2 METHODS

2.1 Gradient Lasso Algorithm

We are interested in investigating the relationship between the survival time and the covariates such as gene expression levels and clinical variables. Let x 1, ..., x m be gene expression levels of m genes and z 1, ..., z p be the clinical variables. Usually, the sample sized denoted by N is smaller than m in microarray data. For i = 1, ..., N, the i-th datum is denoted as (t i, δ i, x i1, ..., x im, z i1, ..., z ip) for i = 1, ..., N where δ i is the censoring indicator and t i denotes the survival time if δ i = 1 or censoring time if δ i = 0. Let x i = (x i1, ..., x im)' be the i-th observation vector of the gene expression level of m genes and z i = (z i1, ..., z ip)' be the i-th observation vector of clinical variables.

We consider the following Cox regression model for the hazard at time t:

\[ \lambda(t) = \lambda_0(t) \exp(x' \beta + z' \gamma), \]

where \( \lambda_0(t) \) is a baseline hazard function, \( \beta = (\beta_0, \beta_1, ..., \beta_m)' \) and \( \gamma = (\gamma_1, ..., \gamma_p)' \) are the coefficient vectors, \( x = (1, x_1, ..., x_m)' \) is the vector of gene expressions, \( z = (z_1, ..., z_p)' \) denotes the vector of clinical variables. Then the partial likelihood (Cox, 1972) is written as

\[ L(\beta, \gamma) = \prod_{i \in R} \exp \left( \frac{1}{\gamma_i} \sum_{j \in R_i} (x_j' \beta + z_j' \gamma) \right), \]

where \( R \) denotes the index set of the events and \( R_i \) is the set of indices of the individuals at risk at time \( t_i \). Let \( l(\beta, \gamma) \) be the log-likelihood.

* to whom correspondence should be addressed
2.2 Gradient LASSO

For a given convex function $C(w)$, $w \in \mathbb{R}^n$, we are to minimize $C(w)$ on $S = \{w : \sum_{j=1}^n |w_j| \leq 1\}$. The gradient LASSO algorithm solves this problem by iteratively updating the solution through the addition and deletions steps. In the addition step, it finds a coordinate at which $C(w)$ decreases most rapidly. To be more specific, let $\nabla C(w)$ be the gradient of $C(w)$ with respect to $w$. The Taylor expansion implies

$$C(w[\alpha, v]) \approx C(w) + \alpha < \nabla C(w), v - w >$$

where $w[\alpha, v] = (1 - \alpha)w + \alpha v$ for $v \in S$. It can be easily shown that

$$\min_{v \in S} < \nabla C(w), v > = \min_{j=1, \ldots, p} \pm \partial C(w)/\partial w_j$$

Hence, the desired direction is a unit vector $v$ such that in the $j$-th coordinate and 0 elsewhere, where

$$j = \arg\min_{j=1, \ldots, p} \{\min[\partial C(w)/\partial w_j, -\partial C(w)/\partial w_j]\}.$$  

After finding $v$, the addition step updates the current solution $w$ by $w[\hat{\alpha}, v]$ where

$$\hat{\alpha} = \arg\min_{\alpha \in [0, 1]} C(w[\alpha, v]).$$

Note that the updated solution always satisfies the constraint $S$ since we take the convex combination. Also, after each addition step, $C(w)$ always decreases since we find $\hat{\alpha}$ so.

In the deletion step, we update the all non-zero solution vector among $w$ simultaneously as follows. For a given current solution $w$, let $\nabla^+ C(w) = (\partial C(w)/\partial w_k I(w_k \neq 0))$. Then, we move $w$ to the direction $-\nabla^+ C(w)$. However, such a move may result in the violation of the constraint $S$. For resolving this problem, we replace $\nabla^+ C(w)$ by $\nabla^+ C(w) = \nabla^+ C(w) - \nabla^+ C(w)\text{sign}(w)$ where

$$\nabla^+ C(w) = \sum_{k=1}^p \nabla^+ C(w)k\text{sign}(w_k)/\sum_{h=1}^p I(w_h \neq 0)$$

and $\text{sign}(w) = (\text{sign}(w_k), k = 1, \ldots, p)$. Here, $\text{sign}(w)$ is 1 if $w > 0$, 0 if $w = 0$ and -1 if $w < 0$. Then, it is easy to see that there exists $\delta > 0$ such that $w - h\nabla^+ C(w) \in S$ for all $h \in [0, \delta]$. Now, we move $w$ toward $\nabla^+ C(w)$ until one of the non-zero coefficients becomes 0. That is, we let $\delta = \min\{w_k/\nabla^+ C(w)k, w_k \neq 0\}$. Finally, we update $w$ by $w + \hat{h}\nabla^+ C(w)$ where

$$\hat{h} = \arg\min_{h \in [0, \delta]} C(w + h\nabla^+ C(w)).$$

Note that $\hat{h}$ is the deleted coefficient. Details of the gradient LASSO algorithm can be found in [7].

2.3 Performance measures

In the microarray-DLBCL survival literature, the predictive performance of a model is typically assessed as follows: (i) risk scores based on the fitted model are computed for patients in a ( withheld) test data set, (ii) strata (usually two) are created based on thresholding these scores, and (iii) log-rank testing of between-strata survival differences is performed. The greater the achieved significance the more predictive the model is deemed. Limitations of this approach include not just the arbitrariness of the imposed stratification but, more importantly, the familiar shortcoming of p-values not necessarily capturing effect size/explained variation. A more refined approach is afforded by the use of time-dependent ROC curves, proposed by [8] and used in the present context by [9]. The time-dependent ROC curves, proposed by [8] and used in the present context by [9], are a more refined approach is afforded by the use of time-dependent ROC curves, proposed by [8] and used in the present context by [9].

1. Initialize: $w = 0$ and $l = 0$.
2. Do until convergence:
   a. Set $l = l + 1$.
   b. Addition:
      (1) Compute the gradient $\nabla C^{(l)}(w) = \{\partial C^{(l)}(w)/\partial w_1, \ldots, \partial C^{(l)}(w)/\partial w_p\}^T$.
      (2) Find the $(j, \gamma)$ which minimizes $\gamma \partial C^{(l)}(w)/\partial w_j$ for $j = 1, \ldots, p, \gamma = \pm 1$.
   c. Deletion:
      (1) Calculate $\nabla^+ C^{(l)}(w)$
      (2) Find $\hat{h}$ where
      $$\hat{h} = \arg\min_{h \in [0, \delta]} C^{(l)}(w + h\nabla^+ C^{(l)}(w)).$$
      (3) Update $w = w + \hat{h}\nabla^+ C^{(l)}(w)$.
3. Return $\beta = \lambda w$.

Fig. 1. The gradient lasso algorithm.

3 RESULTS

3.1 DLBCL Data

This data is taken from [7], consisting of 240 samples from patients with diffuse large B-cell lymphoma (DLBCL), with gene expression measurements for 7399 genes. The outcome was survival time,