Regularization Paths for Stratified Cox's Proportional Hazards Model via Coordinate Descent

Fang Cai

December 14, 2019

1 Abstract

Cox model is prominent for survival analysis. We want to apply cox model for pan-canceer data, with genes as features and survival times as response. The goal is to discover genes that affects survival time of the patient. A very meaningful real life application.Since the data is very high dimensional ($p > n$), we use $L1$ or $L1 + L2$ penalization. The dataset contains genes of patients of different type of cancers and from different hospital. In order to take this in to consideration, we include stratification in the cox model.

There is no ready algorithm for solving a stratified cox model wih elastic penalization. I designed an effective algorithm (fang) and wrote an R package for this. Our algorithm fits via cyclical coordinate descent, and employs warm starts to find a solution along a regularization path. We demonstrate the efficacy of our algorithm on simulated data sets.

2 Introduction

We have data of the form $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ where y_i, the observed time, is a time of failure if δ_i is 1 or right-censoring if $delta_i$ is 0. We further let $t_1 < t_2 < \cdots < t_m$ be the increasing list of unique failure times, and $j(i)$ denote the index of the observation failing at time t_i. We want to study the relationship between predictor variables and survival time. Cox proportional hazards model:

$$h_i(t) = h_0(t)e^{x_i\beta} \tag{1}$$

Log partial likelihood function for cox model:

$$l(\beta) = \sum_{i=1}^{m} \left(x_{j(i)}^T \beta - \log \left(\sum_{j \in R(i)} e^{x_j^T \beta} \right) \right) \tag{2}$$

where $R(i)$ is the set of indices at risk at time t_i.
Note that there is no intercept term in cox model. Cox model with elastic net penalization (scale the log likelihood by a factor of $\frac{2}{n}$ for convenience

$$\frac{2}{n} \left(\sum_{i=1}^{m} x_{j(i)}^T \beta - \log \left(\sum_{j \in R(i)} e^{x_j^T \beta} \right) \right) - \lambda P_\alpha(\beta) \tag{3}$$

2.1 Stratified Cox model

Suppose now we have Stratified data: $(y_i, x_i, \delta_i, z_i)$.

The stratified Cox model allows the form of the underlying hazard function to vary across levels of stratification variables.

$$h_i(t) = h_{z_i}(t)e^{x_i^T \beta}. \tag{4}$$

We can only compare hazards within the same strata, so we have

$$R(i) = \{j : y_j > t_i \text{ and } j \in z(i)\} \tag{5}$$

in (2).

Decompose the log-likelihood function:

$$l(\beta) = \sum_{s=1}^{S} l_s(\beta) l(\eta) = \sum_{s=1}^{S} l_s(\eta_s) \tag{6}$$

where $\eta_s = X_s\beta$ and l_s is the log likelihood function (2) for stratum s, $\{(x_i, y_i, \delta_i) \mid z_i = s\}$. Since the log-likehood can be decomposed into sum of log-likehood within the stratum, corresponding gradient and hessian reduce to the counterpart within the stratum.

$$l'(\eta)_k = l'_s(\eta_s)_k \tag{7}$$

$$l''(\eta)_{kk} = l''_s(\eta_s)_{kk} \tag{8}$$

3 Related work

The idea for adding lasso penalty $\lambda \sum |\beta_j|$ to log partial likelihood to select features was mentioned in Simon et al. (2011) for lasso/Cox model

In this work we employ cyclical coordinate descent. This method has been applied to penalized regression and in particular, elastic net penalties; recently by Friedman et al. (2010), van der Kooij (2007) and Wu and Lange (2008). Friedman et al. (2010) also recognized the strength of employing warm starts to solve the problem along a path of constraint values.

4 Dataset and Features

- Data: feature measurements $x_{ij}, i = 1, \ldots N$ and $j = 1, \ldots p$ for N individuals and p features (genes)
- Censored survival times $(y_i, \delta_i), 1 = 1, 2, \ldots n$ for each individual
- Each individual falls in one of K cancer classes

5 Method

- Modeling: cox proportional hazard model with stratification
- Objective function: Partial likelihood for stratified cox model plus elastic net penalization
- Optimization: Newton-Raphson and coordinate descent for solving Newton-Raphson step. (L1 norm is not differentiable, that's why we use coordinate descent)
- Cross validation: Special method for cross validation since the cox model likelihood function is not additive.

5.1 Algorithm for optimizing it

Objective (cox model with elastic net penalization)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left[\frac{2}{n} \left(\sum_{i=1}^{m} x_{j(i)}^T \beta - \log \left(\sum_{j \in R(i)} e^{x_j^T \beta} \right) \right) - \lambda P_\alpha(\beta) \right] \tag{9}$$

where,

$$\lambda P_\alpha(\beta) = \lambda \left(\alpha \sum_{i=1}^{P} |\beta_i| + \frac{1}{2}(1-\alpha) \sum_{i=1}^{p} \beta_i^2 .n \right) \tag{10}$$

Newton-Raphson A two term Taylor series expansion of $l(\beta)$ at $\tilde{(}\beta)$ has the form

$$l(\beta) \approx l(\tilde{\beta} + (\beta - \tilde{\beta}^T \dot{l}(\tilde{\beta}) + (\beta - \tilde{\beta})^T \ddot{l}(\tilde{\beta}(\beta - \tilde{\beta})/2 \tag{11}$$

$$= l(\tilde{\beta}) + (X\beta - \tilde{\eta})^T l'(\tilde{\eta}) + (X\beta - \tilde{\eta})^T l''(\tilde{\eta})(X\beta - \tilde{\eta})/2 \tag{12}$$

where $\tilde{\eta} = X\tilde{\beta}$. Simple algebra gives us

$$l(\beta) \approx \frac{1}{2}(z(\tilde{\eta}) - X\beta)^T l''(\eta)(z(\tilde{\eta}) - X\beta) + C(\tilde{\eta}, \tilde{\beta}) \tag{13}$$

where

$$z(\tilde{\eta}) = \tilde{\eta} - l''(\tilde{\eta})^{-1} l'(\tilde{\eta}) \tag{14}$$

and $C(\tilde{\eta}, \tilde{\beta})$ does not depend on $\tilde{\beta}$

Speed up In order to speed up the algorithm, we instead replace $l''(\tilde{\eta})$ by a diagonal matrix with the diagonal entries of $l''(\tilde{\eta})$. We denote the ith diagonal entry of $l''(\tilde{\eta})$ by $w(\tilde{\eta})_i$.

$$z(\tilde{\eta})_k = \tilde{\eta} - \frac{l'(\tilde{\eta})_k}{w(\tilde{\eta})_k} \tag{15}$$

Pathwise algorithm:

1. Initialize $\tilde{\beta} = 0$, and set $\tilde{\eta} = X\tilde{\beta} = 0$
2. For $\lambda = \lambda_{\max} > \cdots > \lambda_{\min}$:
   (a) Compute $l''(\tilde{\eta})$ (With function coxgrad from Raillsn Li), $l''(\tilde{\eta})$ (manually from Noah's formula), and $z(\tilde{\eta})$
   (b) Find $\hat{\beta}$ minimizing (glmnet , weighted lasso, family = "gauss", weight = w, standardized = F)

$$\frac{1}{n} \sum_{i=1}^{n} w(\tilde{\eta})_i (z(\tilde{\eta})_i - x_i^T \beta)^2 + \lambda P_\alpha(\beta) \tag{16}$$

   (c) Set $\tilde{\beta} = \hat{\beta}$ and, $\tilde{\eta} = X\hat{\beta}$
   (d) Repeat steps 2a-2c until convergence of $\hat{\beta}$

To solve (16), we use coordinate descent:
For $k = 1, 2, \ldots, p, 1, 2$, minimize (16) in β_k with all $l \neq k$ fixed. The solution is given by

$$\hat{\beta}_k = \frac{S\left(\frac{1}{n}\sum_{i=1}^{n} w(\tilde{\eta})_i x_{ik} \left[z(\tilde{\eta})_i - \sum_{j \neq k} x_{ij}\beta_j \right], \lambda\alpha\right)}{\frac{1}{n}\sum_{i=1}^{p} w(\tilde{\eta})_i x_{ik}^2 + \lambda(1-\alpha)} \tag{17}$$

with

$$S(x, \lambda) = \text{sgn}(x)(|x| - \lambda)_+ \tag{18}$$

$$w(\tilde{\eta})_k = l''(\tilde{\eta})_{k,k} = \sum_{i \in C_k} \left[\frac{e^{\tilde{\eta}_k} \sum_{j \in R(i)} e^{\tilde{\eta}_j} - (e^{\tilde{\eta}_k})^2}{(\sum_{j \in R(i)} e^{\tilde{\eta}_j})^2} \right] \tag{19}$$

$$z(\tilde{\eta})_k = \tilde{\eta}_k - \frac{l'(\tilde{\eta})_k}{l''(\tilde{\eta})_{k,k}} = \tilde{\eta}_k + \frac{1}{w(\tilde{\eta})_k} \left[\delta_k - \sum_{i \in C_k} \left(\frac{e^{\tilde{\eta}_k}}{\sum_{j \in R(i)} e^{\tilde{\eta}_j}} \right) \right] \tag{20}$$

and C_k is the set of i with $t_i < y_k$ (the times for which observation k is still at risk).

6 Experiment

Using simulated data with different baseline hazard, the plots below shows that our method(fang) perform much better at reconstructing the real parameters than the algorithm in the glmnet package in terms of mse and bias.

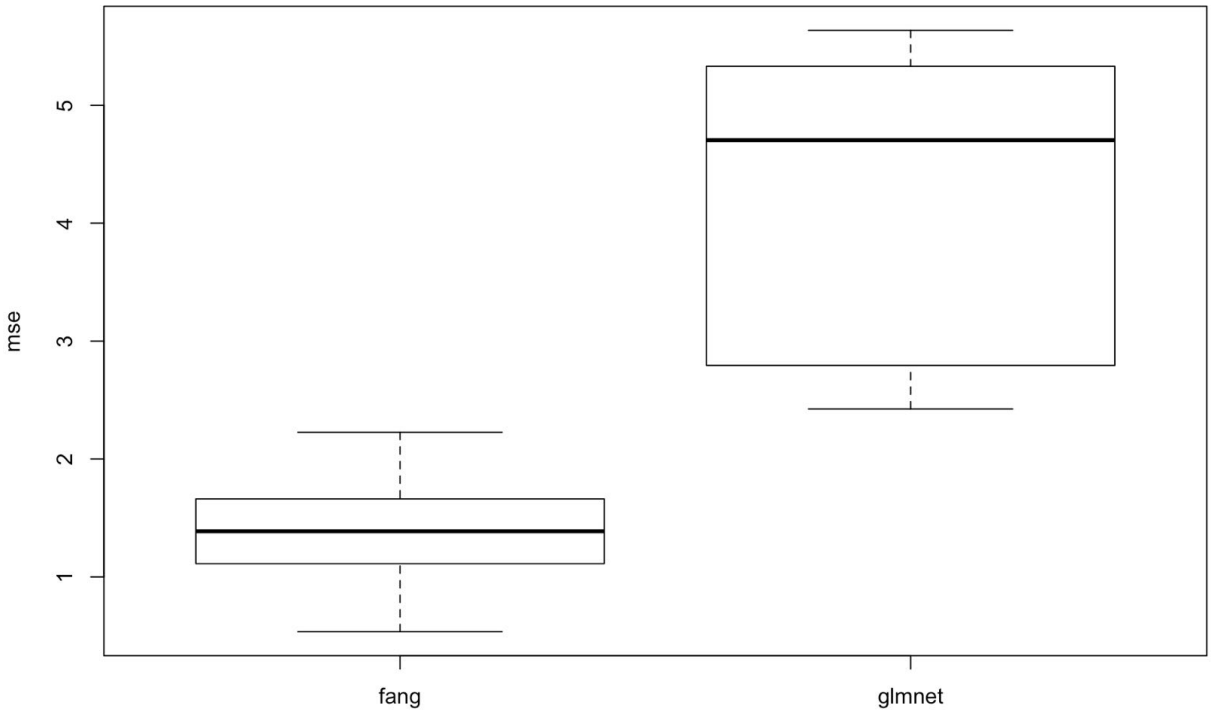

Figure 1. Distribution of mse for 20 simulations with 15 strata, 1000 samples and 20 features with 5 supports

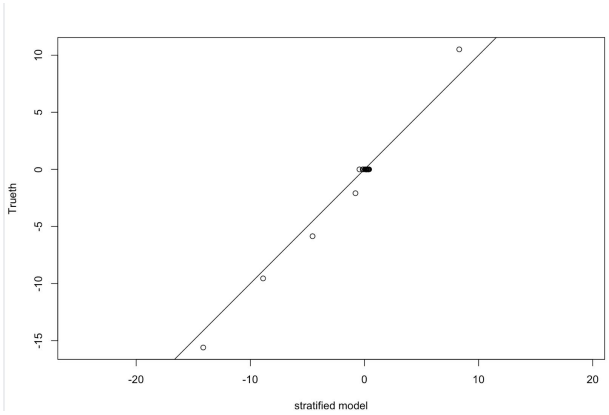

Figure 2. Bias of fang estimated from average 20 simulations with 15 strata, 1000 samples and 20 features with 5 supports

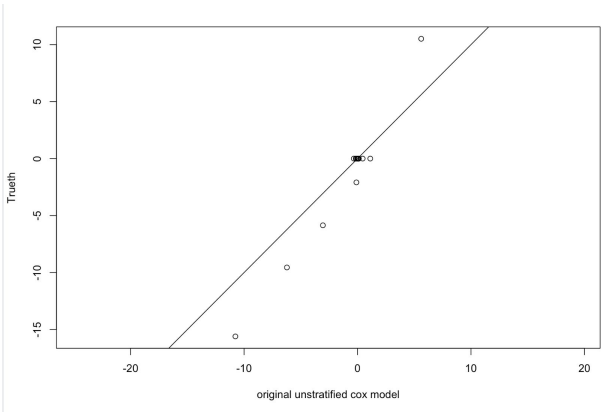

Figure 3. Bias of glmnet estimated from average 20 simulations with 15 strata, 1000 samples and 20 features with 5 supports

7 Conclusion/Future Work

1. Extensions: consider models

   A. $(h_0(t), \beta)$, standard PH model

   B. $(h_{0k}(t), \beta)$ standard stratified model

   C. $(h_0(t), \beta_k)$ main effects for strata

   D. $(h_{0k}(t), \beta_k)$ separate model for each strata

   One could consider blending any of these pairs e.g. $\alpha A + (1-\alpha)B$, using cross-validation to choose α.

   How to blend? Could form linear combination $\tilde{\beta} = \alpha\hat{\beta}_A + (1-\alpha)\hat{\beta}_B$ or instead form objective function $J(\beta) = \alpha log(PL_A(\beta)) + (1-\alpha) log(PL_B(\beta))$ where PL_A, PL_B are partial likelihoods for models A, B, and optimize it.

2. Extensions: add modifiers β_{jk} for feature j, class k with L_1 penalties.

   Alternatively use *reluctant interaction modelling* idea (Gui et al arXiv) to add modifiers β_{jk}

8 Reference

Van der Kooij AJ (2007). Prediction Accuracy and Stability of Regrsssion with Optimal Scaling Transformations. Ph.D. thesis, Department of Data Theory, Leiden University. URL http: //hdl.handle.net/1887/12096.

Wu T, Lange K (2008). "Coordinate Descent Procedures for Lasso Penalized Regression." The Annals of Applied Statistics, 2(1), 224–244.

Friedman J, Hastie T, Hoefling H, Tibshirani R (2007). "Pathwise Coordinate Optimization." The Annals of Applied Statistics, 2(1), 302–332.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software, 33(1), 1–22. URL http: //www.jstatsoft.org/v33/i01/.