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

Fang Cai ffcai@stanford.edu
project supervised by
Robert Tibshirani tibs@stanford.edu

Survival models for pan-cancer data

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

$$h_k(t|x) = h_{0k}(t) \exp(x^T \beta)$$

- Partial likelihood for one stratum

$$L_k(\beta) = \prod_{i \in D_k} \left(\frac{\exp(x_i^T \beta)}{\sum_{j \in R_{ik}} \exp(x_j^T \beta)} \right)$$

where D_k are the indices of deaths in stratum k and R_{ik} are the risk sets in stratum k

- Overall PL

$$L(\beta) = \prod_{k=1}^{K} L_k(\beta)$$

- Idea: add lasso penalty $\lambda \sum |\beta_j|$ to log partial likelihood, to select features

See `http://statweb.stanford.edu/~nsimon/CoxNet.pdf` for lasso/Cox model

- Extensions: consider models **(Future)**

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.

- Extensions: add modifiers β_{jk} for feature j, class k with L_1 penalties. See data shared lasso

`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5650251/` for a related model

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

Abstract

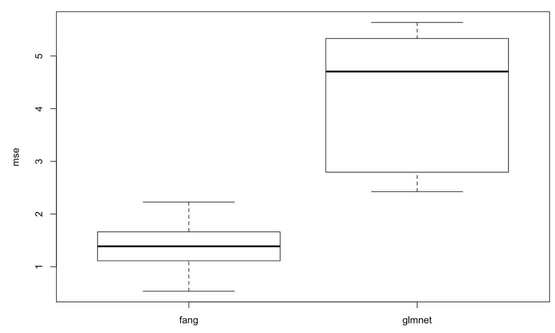
Cox model is important for survival analysis. We want to apply cox model for pan-canceer data, with genesas features and survival times as response. The goal is to discohat affects survival time of thepatient. Since the data is very high dimensional (p > n), we must use L1 or L1 +L2 penalization. The 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. I designed the algorithm and developped a R package for solving a stratified cox model with regularization.

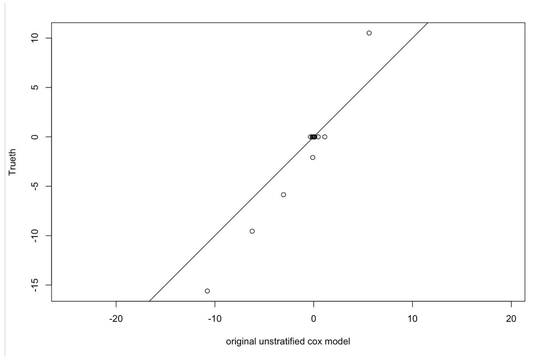
Method

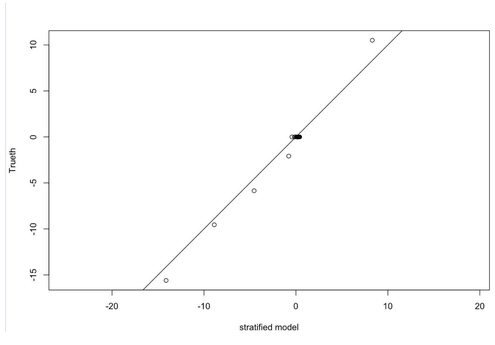
Objective function: Partial likelihood for stratified cox model plus elastic net penalization
Optimazation: Netown-Raphon and coordinate descent for solving Netown-Raphon. (L1 norm is not differentiable, that's why we use coordinate descent)
Cross validation: special method for non-additive objective function.

Discussion

Our method (fang) performs much better than the current unstratified algorithm in glmnet in terms of both mean squared error and bias as shown in the plots below. I hope this project will become a basic tool for high dimensional survival analysis in the future and my algorithm will be added to the glmnet package.







http://www.gytaobao.cn:9292/upload/shipin.mp4