Project Goal: Design a simulation study comparing estimators (eg. Penalized cox, penalized casebase) as well as incorporating other penalization methods (eg. LARS-Cox, SCAD) and exploring various methods of simulation.

Current Goal: to look for a simulation model (a way to simulate data), 3 papers do very different things.

1. Gui and Li (2005), Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.

Paper Goal:

* Use LARS-Cox method to build a parsimonious predictive model used to classify future patients into high/low risk groups based on gene expression profiles and survival times of previous patients.
* To also use the method to select important/relevant genes related to a patient’s survival outcome.
* Compare LARS-Cox method with other methods in predicting survival.

Simulation Study:

* Assume 20 out of 500 genes are related to survival time.
  + 10 beta coefficients generated from U(-1, -0.1)
  + 10 beta coefficients generated from U(0.1, 1)
* Baseline hazard: Weibull(shape = 5, scale = 2)
* Censoring times: U(2, 10)
* Gene expression (X’s): 100x500 dataset from U(-1.5, 1.5)
* 480 genes are not related to survival but may be correlated with 20 relevant genes.
* “maximum possible correlations” (of 0, 0.71, 0.82, 0.87) to show how changes in correlations between relevant and irrelevant genes affect LARS-Cox procedure in selecting relevant genes.
  + How many of the 20 relevant genes are selected out of 500.
* Training and testing data generated in the same way, each with 100 patients.

Pros:

* We know the true Betas, and can use these Betas to generate survival times, such that the generated survival times are related to the X’s.

Cons:

* Generating the dependency between genes is very complicated.

Performance Criteria:

* Performance criterion: time-dependent AUCs.
* For each simulation (100 simulations), n = 100 patients, 500 gene expression levels, max possible gene correlation = 0.82.
  + 4 methods: LARS-Cox, L2, PC-PCR, and SPCA.
  + For each method, build model based on training data and predict risk scores on test data.

1. Wu (2012), Elastic Net for Cox’s Proportional Hazards Model with a Solution Path Algorithm.

Paper Goal:

* Develop a solution path algorithm for the elastic net penalty in Cox’s PH model.

Simulation Study:

* Example goal: show that LASSO solution path is not piecewise linear.
* Generate 3x40 “X” matrix of predictors N(0, 0.9)
* Generate true Betas (2, -2, 2.5)
* Set baseline hazard = 1
* Generate censoring times U(0, 8)
* Generate survival times from the hazard:
  + -log(U(0,1)) / exp(X\*Beta)

Pros:

* We know the true Betas, and can use these Betas to generate survival times, such that the generated survival times are related to the X’s.

Performance Criteria:

* Performance criterion: figure showing the LASSO solution path to not be piecewise linear.

1. Biar and Tibshirani (2004), Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data.

Paper Goal:

* Identify cancer subtypes (clusters) through a procedure that combines gene expression data and clinical data.
* Train a classifier (use semi-supervised learning techniques) that can diagnose which type of cancer a future patient has.

Simulation Study:

* Generate 5000 genes and 100 samples for training data, testing data generated the same.
* Generate “X” gene expressions as standard normal N(0, 1).
* Generate survival times from normal distribution.
* Generate censoring times N(10, 3).
  + If censoring time < survival time, the observation is censored.
* Applied/compared classifier methods to simulated data.

Cons:

* The survival times are not related with the X’s, since the true values of Betas are not specified.

Performance Criteria:

* Performance criterion: # of misclassified samples, CV errors, test errors, R^2.