**Reference:**

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

**Introduction and Methods**

Motivation:

* Lots of research previously done on predicting cancer class using gene expression data
* there has been less research in linking gene expression profiles to censored survival data
* we want to build models that can do this with good accuracy and parsimony (without spending money)
* censored survival phenotypes vs. categorical cancer phenotypes
  + censored survival phenotypes (e.g. time to cancer/death) is more informative due to large variability in time
* Cox regression is popular for censored survival data
  + BUT due to high dimensional gene expression data: the standard maximum Cox partial likelihood method cannot be applied directly to obtain the parameter estimates
  + AND high collinearity problem: expression levels of genes are often highly correlated
* Solution to high collinearity: (variable selection)
  + L2 variable selection
    - BUT it uses all genes in the prediction (cannot select relevant genes)
  + L1 variable selection (selects relevant genes)
    - BUT needs to use a quadratic programming procedure which cannot be applied directly to settings when sample size (n) is much smaller than # of predictors (high-dimensional)
  + LARS variable selection (selects predictors by its current correlation/angle with the response, correlation between predictor and current residuals)
    - LARS algorithm can perform variable selection in high-dimension and low-sample settings

Method:

* use LARS-Cox procedure to select relevant genes and build a predictive model in high-dimensional (many predictors) and low-sample (small n) settings
* predictive model classifies patients into high-risk or low-risk group
* use the LARS algorithm to obtain solutions for Cox model with L1 penalty in the high-dimensional setting

LARS:

Forward stepwise regression

- (1) begin with null model (model with no variables)

- (2) add the most significant variable, one at a time

- fit p simple linear regression models (1 variable and intercept), choose the one with lowest RSS

- then search through remaining p-1 variables to find out which variable should be added to current model to improve RSS

- "most significant" criteria based on:

- smallest p-value

- highest increase in R^2

- highest decrease in RSS

- (3) repeat until some kind of stopping rule is applied

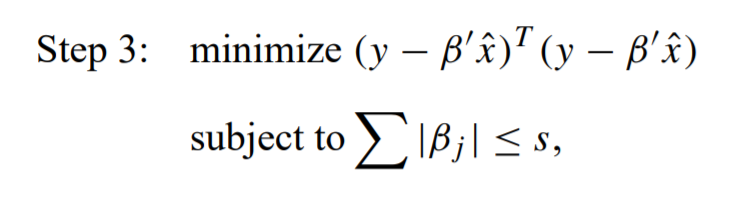
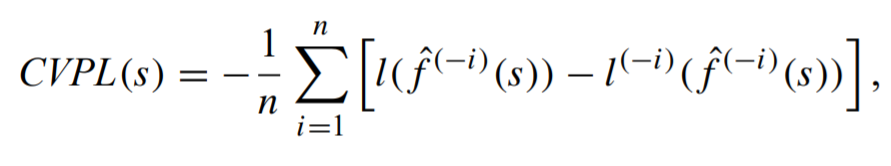
- “stopping rule” determine by: when you add an additional variable to the model, it exceeds the p-value/AIC/BIC threshold of your model, then don’t add the variable and stop

- (4, optional) select the best model from model with 1 predictor, model with 2 predictors, ... model with p predictors

Forward stagewise…

LARS…

LARS-Cox Procedure

* solve a minimization problem using the LARS algorithm subject to a Lasso constraint for a given “s”
* use LARS algorithm to solve
  + 
* choose “s” value that minimizes cross-validated partial likelihood (CVPL)
  + 

Evaluating Predictive Performance

* AUC-ROC curve (performance measurement for classification problems)
* The higher the AUC value at a given time, the better the predictive performance of model
* Time dependent/sensitive AUCs (conditional probability given the censoring state)
  + Important to know
  + How to compute them

**Compare Methods by Simulation Study**

* Compare on:
  + If important covariates can be selected
  + How well the model predicts survival

Effects of between-gene correlations on identifying relevant genes

* For each maximum possible correlation (corr between 20 relevant genes and 480 irrelevant genes) of 0, 0.71, 0.82, and 0.87, 100 datasets (replications) were generated with n=100 individuals
* For each replication, a LARS-Cox model is built to select 20 genes by setting an appropriate “s” value.
  + Table 1 shows the frequencies out of 100 that the 20 relevant “Beta” genes selected by LARS-Cox
  + Observations:
    - Predictors with larger coefficients more likely to be selected
    - As maximum possible correlation increases, the chance of relevant genes with smaller coefficients being selected decreases.
      * Because at each step, LARS-Cox selects only the gene with largest absolute correlation (aka it’s more apparent which gene to select if you increase max corr)
    - As sample size increases (to 200 for max corr = 0.85), more relevant genes are selected.
* In summary, decreased max possible correlation and increased sample size = more relevant genes selected.

Predictive performance and comparison with other methods

* 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 LARS-Cox, used CVPL to choose “s”
  + For each method, build model based on training data and predict risk scores on test data.
  + Criterion for predictive performance: time-dependent AUC
  + Figure 1: shows average AUC curves for 4 methods
* Alternative method: divide patients in test data into high-risk or low-risk groups based on positive or negative predictive risk scores (?)
* In summary, LARS-Cox had a better predictive performance than other methods.

Simulation Study Summary

* LARS-Cox can select genes related to censored phenotypes (e.g. time to event), especially genes with strong effect (high beta).
* Genes with small effect harder to select (small beta), especially if gene correlations are high.
* When gene correlations are high, CVPL tends to select a greater # of genes.

**Application to Prediction of Survival Time of Patients with DLBCL**

Selection of genes related to risk of death

* Table 2: shows the top 10 genes selected by LARS-Cox (as tuning param increases, more genes are selected)

Evaluation of predictive performance

* LARS-Cox method in DLBCL
  + Minimum CVPL was obtained when s=0.28, chose the most parsimonious model
  + Obtained estimates of 4 selected genes, 4 coefficients, all negative, so interpretation..
  + Estimated risk scores for 80 test patients, and got time-dependant AUCs
  + Using zero as a cut-off point for risk scores, patients were divided into 2 groups: positive risk scores and negative risk-scores
    - Figure 2(b): shows KM curves, high-risk group and low-risk group were significantly different
    - Be care of post-inference problem
  + Comparing with the other 3 methods, LARS-Cox showed the most significant difference in risk in KM Curves
    - And LARS-Cox had higher AUC curve comparably
* Similar results were found doing the same thing with lymphoma dataset

**Discussions and Conclusions**

* Importance: very important to predict time to cancer/death after treatment using gene expression profiles prior to treatment
* LARS-Cox was used to:
  + identify important genes that predict survival
  + and build a parsimonious model to predict survival
* LARS-Cox method was found to perform better than L2, PC-PCR, and SPCA in predicting survival
* Future direction:
  + A more comprehensive comparison of different methods (e.g. partial least squares, principal components Cox regression)
* Limitation of LARS-Cox
  + No limitations in terms of # of genes to be used for the model, LARS-Cox can select n-1 genes, (n is sample size)
  + BUT when # of predictors in model is close to sample size n, there is a risk of over-fitting
  + The # of genes selected cannot be more than the sample size.
    - Because as “s” increases (less constraint), and the # of selected genes gets close to # of observations, Lasso may not have a unique solution (aka only 1 set of genes to select)
  + LARS-Cox tends to select only ONE gene from a group of highly correlated genes
    - Is a problem in identifying important and relevant genes
    - Is not a problem if goal is to build model with good predictive accuracy (simple models are preferable)
    - One possible solution: at each LARS variable selection step, select not one gene with largest absolute current inner product but group of genes with similar current inner products
      * Another alternative: use the elastic net penalty (selects more genes because not lasso)
  + LARS-Cox assumes Cox PH assumption, but this may not hold for gene expression data
    - Solutions:
      * Develop robust procedures under mis-specified PH models
      * Use model checking techniques to check assumptions
      * Consider L1 penalized estimation for accelerated failure time models or more general semi-parametric transformation models

Summary

* LARS-Cox method can be useful to build a parsimonious predictive model that can predict survival times accurately (AUC) based on gene expression data, and then survival times/risk scores to be classified into high/low risk groups. LARS-Cox method also useful to select important genes related to patient’s survival.