COM S 573: Home work 3 Spring 2016

Write your name on each page. Maximum score is 30 points, due date is **Wednesday, April 13, 2016** . Please hand in the solutions (CLEAN version) on the due date in class (**hard copy**). Also paste the results of your R code and the code itself into your homework. Make sure your homework is stapled!

1. (a) [4 points] Suppose that $Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$ where $e_1, \ldots, e_n$ are i.i.d. distributed from a $N(0, \sigma_e^2)$. Write out the likelihood for the data and show that it is equivalently to using ordinary least squares.

   (b) [4 points] Assume the following prior for $\boldsymbol{\beta}$: $\beta_1, \ldots, \beta_p$ are i.i.d. according to a Laplace distribution with mean zero and common scale parameter c i.e., $h(\boldsymbol{\beta}) = \frac{1}{2c}\exp(-|\boldsymbol{\beta}|/c)$. You can assume that $Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$ with $e_1, \ldots, e_n$ are i.i.d. distributed from a $N(0, \sigma_e^2)$. Write out the posterior for β in this setting. Argue that the LASSO estimate is the mode for $\boldsymbol{\beta}$ i.e, the most likely value for $\boldsymbol{\beta}$, under this posterior distribution. Determine the value for the parameter λ in the LASSO cost function.

2. (a) [4 points] Suppose we estimate some statistic (e.g. median) based on a sample X. Carefully describe how you might estimate the standard deviation of the statistic. You can make a sketch of the process.

   (b) [2 points] Write an R code that calculates the standard deviation of the median given a sample X.

   (c) [4 points] Suppose you were interested in a $100(1-\alpha)\%$ (pointwise) confidence interval for the correlation coefficient of a sample X and Y (the joint distribution of X and Y is NOT bivariate normal). Clearly explain and derive how you would do this? Write an R code that calculates 95% confidence interval for the correlation coefficient in case of the lawstat data (lawstat.dat on Blackboard).

3. (a) [5 points] Try out some of the regression methods explored in Chapter 6 of the textbook on the Boston Housing data set (available from the MASS library), such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

   (b) [5 points] Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Clearly explain what you will do.

4. [2 points] Describe how you can efficiently solve the LS linear system $(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\mathbf{Y}$ (i.e., by not calculating an inverse) where $\mathbf{X} \in \mathbb{R}^{n\times p}$ has p linearly independent columns, $\beta \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^{n\times 1}$? *Hint*: Think in terms of matrix decompositions (it's not SVD!). Use Wikipedia.