High-Dimensional Data Analysis (BIOS 7600)
Breheny

Assignment 7
Due: May 11

1. *Semi-penalized inference*: Problem is described on slides 36 and 37 of the April 4 (4-4) notes.

2. *Inference for the WHO-ARI data*: In class, we discussed several ideas for carrying out inference (false inclusion rates, sample splitting, bootstrapping, selective inference). Choose two of those ideas and apply them to the World Health Organization study of acute respiratory illnesses (we have analyzed this data a few times already, among them Problem 1.11 from the text).

   In terms of a finished product, provide (a) a paragraph describing the methods you used and how you implemented them (e.g., if there were any tuning parameters, how did you choose them; did you use an `R` package; if so, which one); (b) a summary of results for each approach; (c) a paragraph (or more) commenting on these results and making at least one interesting comparison between the approaches.

3. Let $L(\boldsymbol{\beta})$ denote a differentiable loss function. Consider taking a second-order Taylor series expansion of L about $\tilde{\boldsymbol{\eta}}$, where $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\eta}} = \mathbf{X}\widetilde{\boldsymbol{\beta}}$ (L can be thought of equivalently as a function of $\boldsymbol{\beta}$ or a function of $\boldsymbol{\eta}$). Let $\mathbf{v}$ and $\mathbf{A}$ denote the first and second derivatives of L with respect to $\boldsymbol{\eta}$, and let $\mathbf{z} = \tilde{\boldsymbol{\eta}} - \mathbf{A}^{-1}\mathbf{v}$. Show that, up to a constant,

$$L(\boldsymbol{\beta}) \approx \frac{1}{2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{A}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}).$$

4. Carry out a penalized logistic regression analysis of the case-control prostate cancer study `Singh2002` (you previously analyzed this data using t-tests, back in Assignment 2). Choose two penalties that we have discussed in class and analyze the data using those penalties. Similar to Problem 2, provide (a) a paragraph describing the methods you used, (b) a summary of each method in terms of the number of features selected and predictive accuracy, and (c) comment on at least one interesting difference between the approaches. This could be a global comparison (e.g., method A selects more features than method B) or a comparison of a specific estimated parameter.

5. *Group lasso simulation*: Design a simulation to illustrate the potential advantages of the group lasso in situations where the grouping variable contains useful information. As an end result, try to produce a figure with "grouping" on the horizontal axis and mean squared error on the vertical axis; the figure should show that as "grouping" increases, group lasso outperforms the ordinary lasso by an increasingly large margin. Part of the challenge of this exercise is coming up with some definition of what "grouping" means; be creative!