Kimberly Boucher
CM 451
October 31, 2013
Thesis

Initial Simulations

We assess the effectiveness of CLS as a variable screening method in very high-dimensional problems by using numerical simulations in R to test that CLS correctly identifies predictors on which the response most strongly depends but does not mistakenly identify predictors whose relationship is only noise. Initially we consider only linear relationships, which can be written in the form $y = X \cdot \beta + e$, where y is the vector of observations of the response, X is a matrix of observations whose columns correspond to the predictors, β is a vector of coefficients in decreasing order of magnitude that determine the dependence of the response on each predictor, and e is vector of error terms in each observation, the part of y that is unexplained by its relationship with any of the predictors. We do this in four scenarios: one in which all of the predictors are independent of each other and the response, one in which some of the predictors are dependent on each other but all are independent of the response, one in which all of the predictors are independent of each other and the response depends on some of them, and one in which some of the predictors are dependent on each other and some have a relationship to the response. For each of these scenarios, we run 500 trials with 1000 predictors and 60 observations. For the first scenario, we generate a matrix of all observations of all variables by using the rnorm function to simulate a situation in which all observations are independent of each other, and generate the response the same way. For the second scenario, we use a variance-covariance matrix of all 0.6's except for 1's on the diagonal and generate the observation data by using a multivariate normal distribution (rmvnorm from the library mvtnorm) and generate the response data from an independent standard normal distribution using the rnorm function. For the third and fourth scenarios, we generate the observations of the predictors in the same way as we did for the first and second respectively, but we generate the response data by multiplying the predictor data matrix X by the vector $\beta \neq 0$ of coefficients and then add the vector e generated using rnorm.