Deterministic models and optimization 2017.

Assignment 2: Convexity. Due date: December 7, 2017

Problem 1 [Recovering the correct support] Suppose the data vector y comes from a linear model $y = \mathbf{X}\beta^* + \epsilon$, where only a certain subset of the entries β^* are nonzero. We say that our procedure correctly identifies the support β^* if the estimator $\widehat{\beta}$ has exactly the same nonzero entries as β^*. In this exercise we will study how well the lasso estimator performs, which I believe generated some interest in class. In practice we do not know β^* and we only see $\mathbf{X}$ and y. The aim is to focus on a simplistic scenario, generate data for some known β^*, and see how the lasso estimator performs in recovering the true support.

Consider the two-dimensional case $\beta^* = (a, 0)$ (the second parameter is zero). Let n be the sample size. Generate $\mathbf{X}$ from a two dimensional Gaussian distribution with covariance matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Here $\rho \in (0, 1)$ will measure collinearity of the data. Not to introduce more parameters sample n ϵ's from $N(0, 1)$. Now y is given by $y = \mathbf{X}\beta^* + \epsilon$. Based on this y and $\mathbf{X}$ use some standard lasso procedure (with default options) in R to estimate β^*, e.g. package `glmnet`. Check if the support of this estimator is equal to the support of β^*. Now do, say, 100 replicates, each time checking if the support is correct. Report the probability of recovering the correct support. This of course will highly depend on the parameters n, a, and ρ. Can you guess how? Do your simulations confirm your intuition?

A solution to this problem should contain simulation results for different values of the three parameters together with some discussion of the obtained results.