CDA Spring 2026

Sneha Das

February 16, 2026

Week 3: SPARSE REGRESSION

1. Apply Least angle regression and selection (LARS) for the $p >> n$ sand data set ($\mathbf{X}$: data matrix with 59 observations and 2016 features, $\mathbf{y}$: the measured moisture content in percent for each sand sample). Find a suitable solution using:
   (a) The Cp statistic. Consider whether the C_p-statistic makes sense in this case ($p > n$). Why? Why not?
   - Hint: What happens to your estimate of the noise in the data?

   (b) Using Cross-validation. Remember to center $\mathbf{y}$ and normalize $\mathbf{X}$, but do it inside the cross validation!

2. Find an elastic net solution for the sand data, with suitable choices of regression parameters using cross validation.
   (a) Use the coordinate descent algorithm.
   - Python: Use Python's `linear_model.ElasticNet`.

   (b) Investigate how different values of α affects the number of nonzero parameters in the coordinate descent algorithms.

   (c) What are the pros and cons of the coordinate descent algorithm compared to using LARS?

3. Perform univariate feature selection for the sand data using:
   (a) Bonferroni correction to control the family-wise error rate (FWER). Use $FWER = 0.05$.

   (b) Benjamini-Hochberg's algorithm for FDR. Use an acceptable fraction of mistakes, $q = 0.15$.

   Compare the solutions in terms of number of selected features and selected features. Hint: See the resources for implementations of Benjamini-Hochberg's algorithm.