Tamar Feldman

Biol419, Winter ‘16

Project Proposal

1. My datasets are from the World Bank and the Global Health Observatory data repository of the World Health Organization. I am using datasets on sanitation, population density, healthcare infrastructure and workforce, literacy, and neonatal mortality.
2. I am investigating a series of social, economic, and environmental indicators that potentially affect a country’s health. I will use neonatal mortality as a measure of health. My questions are:
3. Does the variability in the indicators reflect the variability in health across countries?
4. Are there countries whose health is not well explained by the PCA analysis and, if so, why?
5. What combination of these indicators is the best predictor of a country’s health?
6. Identify the most recent year for which there is the most complete data in every indicator. (possibly average values over a few years or do the analysis for a few different time points)
7. Perform descriptive analysis on each indicator. If there is an indicator with low variability, exclude it because it will not help explain variability in health.
8. Perform PCA on the indicators. Determine if the underlying variability reflects the variability in health (similar to the PCA seed type example) by doing cross validated accuracy. Are there countries that have unusually good or bad health?
9. Do LDA using the indicator data. Find the best indicators by cross validated accuracy and then compare with LDA using a combination of these best indicators.
10. Histogram of neonatal mortality in the chosen year(s)
11. Present the descriptive analysis of each indicator in a table
12. Scatterplot of the data along the first 2 or 3 principal components
13. Stairs plot of the eigenvalues
14. Plots of neonatal mortality vs best indicators from LDA
15. Present the cross validated accuracies for the different models in a table