Matrix Calculus and Optimization
Statistics 202B
Professor: Mark S. Handcock

Homework 3

Due Wednesday, February 10, 2016

Note: **MMST** stands for the Alan J. Izenman "Modern Multivariate Statistical Techniques" (2008) online text.

1) Consider the data on the chemical composition of tobacco considered in Section 6.3.3 of **MMST**. The data is in the `MMST` package on `CRAN`. To obtain it, use:

```
install.packages("MMST")
library(MMST)
data(tobacco)
help(tobacco)
head(tobacco)
```

The data are taken from a study on the chemical composition of tobacco leaf samples (Anderson and Bancroft, 1952, p. 205). There are $n = 25$ observations on $m = 6$ covariates variables, percent nitrogen, percent chlorine, percent potassium, percent phosphorus, percent calcium, and percent magnesium. There are three outcome variables of interest (rate of cigarette burn in inches per 1,000 seconds, percent sugar in the leaf, and percent nicotine in the leaf.

Consider a regression model for the rate of cigarette burn in inches per 1,000 seconds based on the $m = 6$ covariates variables. Fit a total least squares model to the data using the `tls()` function. Briefly describe the difference in result from the OLS fit to the same data.

2) Consider the following operation of *data augmentation*. Suppose all input and output variables have been standardized (i.e, centered on their mean and divided by their standard deviation). Augment the $(n \times r)-$ matrix X with r additional rows of the form $H_k = \sqrt{k} I_r$, where k is given and I_r is the $r-$ dimensional identity matrix. Denote the resulting $((n+r) \times r)-$ matrix by X^*. Augment the $n-$ vector Y using r 0s, and denote the resulting $(n+r)-$ vector by Y^*.

Show that the ridge estimator can be obtained by applying OLS to the regression of Y^* on X^*.

Thus, one can carry out ridge regression using standard OLS regression software and obtain the correct ridge estimator. However, much of the rest of the regression output will be inappropriate for the original data (X, Y).

3) Consider the PET Yarn data considered in Section 5.6.1 of **MMST**. Fit a ridge regression model to the centered data.

Plot the ridge trace of the first 60 ridge estimates of the 268 regression coefficients for $\kappa \leq 1$.

Briefly describe how the estimated coefficients are influenced by the value of κ.

4) For the PET Yarn data considered in the previous question, estimate the ridge parameter using cross-validation. **MMST**Section 5.4.2 provides good background. **MMST**Table 5.7 provides a detailed algorithm.
There is also the "One Standard Error Rule" for cross-validation, where you choose the ridge parameter one standard deviation higher than the minimum cross-validation error. Try this rule also. Is one rule better than the other?

5) Consider question 7.7 of **MMST**. The file `SwissBankNotes.txt` consists of six variables measured on 200 old Swiss 1,000-franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height of the bank note, measured on the right, distance of inner frame to the lower border, distance of inner frame to the upper border, and length of the diagonal. Carry out a PCA of the 100 genuine bank notes. Repeat for the 100 counterfeit bank notes, and then for all 200 bank notes combined.
Do you notice any differences in the results?
How would you tell someone what the differences between the genuine and counterfeit notes are?
To read in the data you can use:

```
read.table("http://www.stat.ucla.edu/~handcock/202B/datasets/SwissBankNotes.txt",
skip=22,header=TRUE)
```