Math 420, Spring 2016
Second Solo Homework:
Introduction to the Threads

Due Tuesday, 9 February, 2016

Exercise 1. Compute m_i, v_{ij}, and c_{ij} for each of the following groups of assets based on daily closing price data with uniform weights:

(i) Apple, Google, Exxon-Mobil, UPS, GE, and Ford stock in 2014;
(ii) Apple, Google, Exxon-Mobil, UPS, GE, and Ford stock in 2015;
(iii) Vanguard Total Bond, Vanguard 500, and Russell 1000 and 2000 index funds in 2014;
(iv) Vanguard Total Bond, Vanguard 500, and Russell 1000 and 2000 index funds in 2015.

a. Display m_i as a 6-vector and v_{ij} and c_{ij} as 6×6-matrices for (i) and (ii). Explain the differences between these objects for groups (i) and (ii).
b. Display m_i as a 4-vector and v_{ij} and c_{ij} as 4×4-matrices for (iii) and (iv). Explain the differences between these objects for groups (iii) and (iv).
c. Give explanations for the values of c_{ij} you computed.

Exercise 2. In this problem you will analyze a collection of handwritten digits collected by USPS, see, e.g., the data at the page of Prof. S. Roweis. The origin of this problem comes from a series of experiments in the early 90s which were aimed at developing methods for automated handwriting recognition. One of the main players in this was the United States Postal Service. USPS, in collaboration with Center of Excellence in Document Analysis and Recognition (SUNY Buffalo), collected a large number of addresses and developed a number of databases. You can find a sample of this data here:

http://www.cs.nyu.edu/∼roweis/data/usps_all.mat

This dataset is formed by 1100 examples of images of each of 10 digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) in the 8-bit grayscale format, and they all are derived from a 16×16 image structure. In this database the original images are replaced with vectors by reading the coefficients of each square image "column-wise." As such the whole database represents 11,000 vectors in 256 dimensions (recall that $16^2 = 256$).

a. Download the dataset of handwritten digits collected by USPS.
b. Present as images the following digits from this dataset: 1st 0, 13th 2, 100th 5, and 111th 8.
c. Use $\|\cdot\|_p$ norms with $p = 1, 2, 4, \infty$ to compute the diameters of the following sets:
   (i) set of first 100 1s,
   (ii) set of first 100 9s.

Record computing times that these tasks required, and make an estimation of times needed to compute diameters of the whole sets of 1s and 9s, resp.

Next, compute those diameters for the whole sets of 1100 digits, record the actual times and compare to your estimates. (Please note that depending on your computer this can be a time-consuming exercise.)

Was your prediction accurate? If not, describe what do you think went wrong?

Analyze the results for different norms. Do the diameters change significantly when you change p and, if so, how?