Mathematical Introduction to Data Science **Sep 15, 2012**

Project I

Instructor: Yuan Yao *Due: Monday Oct 8, 2012*

The first project aims to have more experiments on PCA/MDS. Team work is encouraged, but you have to *write your proposal yourself* with a full citation of partners' work. Submit to TA your report with your source codes (as appendix in report or zipped files).

On course website, you will find the following datasets

1. *Journey to the West:* This dataset contains a 302-by-408 matrix X with 302 characters and 408 scenes collected from one hundred chapters in the classic novel by WU, Cheng-En. Each element contains either 0 or 1 indicating if the character appears in the scene.

   `http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/xiyouji/xiyouji.mat`

   The construction of this matrix from original data can be found in Matlab file:

   `http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/xiyouji/readData.m`

2. *Hand-written Digits:* The website

   `http://www-stat.stanford.edu/~{}tibs/ElemStatLearn/datasets/zip.digits/`

   contains images of 10 handwritten digits ('0',...,'9');

3. *S&P500 Prices:* This dataset contains a data matrix $X \in \mathcal{R}^{n\times p}$ of $n = 1258$ consecutive observation days and $p = 452$ daily closing stock prices, and the cell variable "stock" collects the names, codes, and the affiliated industrial sectors of the 452 stocks.

   `http://www.math.pku.edu.cn/teachers/yaoy/data/snp452-data.mat`

   contains images of 10 handwritten digits ('0',...,'9');

4. *SNPs of World-wide Populations:* This dataset contains a data matrix $X \in \mathcal{R}^{p\times n}$ of about $n = 650,000$ columns of SNPs (Single Nucleid Polymorphisms) and $p = 1064$ rows of peoples around the world. Each element is of three choices, 0 (for 'AA'), 1 (for 'AC'), 2 (for 'CC'), and some missing values marked by 9.

   `http://www.math.pku.edu.cn/teachers/yaoy/data/ceph_hgdp_minor_code_XNA.txt.zip`

   Moreover, the following file contains the region where each people comes from, as well as two variables `ind1` and`ind2` such that $X(\texttt{ind1}, \texttt{ind2})$ removes all missing values.

   `http://www.math.pku.edu.cn/teachers/yaoy/data/HGDP_region.mat`

   Some results by PCA can be found in the following paper, Supplementary Information.

   `http://www.sciencemag.org/content/319/5866/1100.abstract`

Pick up your favorite dataset(s), perform PCA or MDS to analyze and visualize the data, with denoising and outliers removal if necessary. Attention: the last dataset is relatively big with about 2GB size.