Mathematical Introduction to Data Science **Nov 5, 2012**

Mini-Project II

Instructor: Yuan Yao *Due: Monday Nov 19, 2012*

This project aims to exercise the tools in the class, such as random projections, robust PCA, sparse PCA, and MDS with uncertainty, etc., based on the real datasets. In the below, we list 6 candidate datasets for your reference.

1. Choose the dataset in your favorite. In your report, raise a problem that you are interested to attack with the tools you learned. The following problems are for your examples:

   Robust PCA will decompose the data matrices into two parts, a low-rank and a sparse, which can be applied to all the datasets below. What's the meaning of sparse part and what's the meaning of low-rank part in your data?

   Sparse PCA will find PCA whose support are only subsets of the variables. Journey to the West and Chinese Medicine datasets are probably suitable for such a tool, as we would like to find small teams of characters dominating the story, or small sets of essential herb curing most of similar deseases (flu).

   Random Projection is to do dimensionality reduction if one wants to find SVD of X of large dimensionality. This particularly suits the analysis of SNPs data with $650k$ sites, where you can random projects the matrix into a subset of SNPs and see if top eigenvectors are still approximately correct.

   MDS with uncertainty is to reconstruct Euclidean coordinates when pairwise distances are partially observed and contaminated. Can one reconstruct Euclidean coordinates of Bird Flu virus from geodesic distances and sequence distances, respectively, and compare one against another?

2. Try to apply the methods to analyze your data. Analyze the phenomena you observed and explain whether or not the results answer your problems in the first step and *why*. Finally in your report list the kind of *open problems* in your project and possible future directions in your mind.

Datasets

1. *Bird Flu Dataset:* (courtesy of Steve Smale and Cissy) This dataset 162 H5N1 (bird flu) virus sequences discovered around the world:

   `http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_seq162.txt`

   Locations of such virus discovered are reported with latitude and longitude coordinates on the globe:

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

Pairwise geodesic distances between these 162 sites are constructed as

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

A kernel-induced l_2-distances between 162 virus sequences are given in

`http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_l2dist.txt`

2. *Chinese Medicine Dataset:* (courtesy of Zhi Geng) This dataset contains a 2885-by-965 matrix X with 2885 diseases (mostly flu) and 965 Chinese herbal prescription drugs collected from traditional Chinese medicine literature. Each element contains either 0 or 1 indicating if the herbal drug is used for the disease. Variable `desease2885` is a 1-by-2885 cell collecting the description of deseases and `herb965` contains the names of those herbal drugs.

   `http://www.math.pku.edu.cn/teachers/yaoy/data/cmed965.mat`

3. *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/data/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/data/xiyouji/readData.m`

4. *Hand-written Digits:* The website

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

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

5. *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’);

6. *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(`ind1`, `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`

   Attention: this last dataset is relatively big with about 2GB size.