**A Mathematical Introduction to Data Analysis** **29 Sep, 2017**

# Mini-Project 1

*Instructor: Yuan Yao* *Due: 0:00am Friday 13 Oct, 2017*

## 1 Mini-Project Requirement and Datasets

This project as a warmup aims to exercise the tools in the class, such as PCA/MDS and their various extensions, biased estimators, etc., based on the real datasets. In the below, we list some candidate datasets for your reference.

1. Pick up ONE (or more if you like) favorite dataset below to work. If you would like to work on a different problem outside the candidates we proposed, please email course instructor about your proposal.

2. Team work: we encourage you to form small team, up to THREE persons per group, to work on the same problem. Each team just submit ONE report, *with a clear remark on each person's contribution*. The report can be in the format of a *technical report within 8 pages*, e.g. NIPS conference style

   `https://nips.cc/Conferences/2016/PaperInformation/StyleFiles`

   or of a *poster*, e.g.

   `https://github.com/yuany-pku/2017_math6380/blob/master/project1/DongLoXia_poster.pptx`

3. In the report, (1) design or raise your scientific problems (a good problem is sometimes more important than solving it); (2) show your main results with a careful analysis supporting the results toward answering your problems. Remember: scientific analysis and reasoning are more important than merely the performance results. Source codes may be submitted through email as a zip file, or as an appendix if it is not large.

4. Submit your report by email or paper version no later than the deadline, to the following address (datascience.hw@gmail.com) with Title: CSIC 5011: Project 1.

## 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 Finance Data

The following data contains 1258-by-452 matrix with closed prices of 452 stocks in SNP'500 for workdays in 4 years.

`http://math.stanford.edu/~yuany/course/data/snp452-data.mat`

or in R:

`http://math.stanford.edu/~yuany/course/data/snp500.Rda`

# 4 Animal Sleeping Data

The following data contains animal sleeping hours together with other features:

`http://math.stanford.edu/~yuany/course/data/sleep1.csv`

# 5 US Crime Data

The following data contains crime rates in 59 US cities during 1970-1992:

`http://math.stanford.edu/~yuany/course/data/crime.zip`

Some students in previous classes study crime prediction in comparison with MLE and James-Stein, for example, see

`https://github.com/yuany-pku/2017_math6380/blob/master/project1/DongLoXia_slides.pptx`

# 6 NIPS paper datasets

NIPS is one of the major machine learning conferences. The following datasets collect NIPS papers:

## 6.1 NIPS papers (1987-2016)

The following website:

`https://www.kaggle.com/benhamner/nips-papers`

collects titles, authors, abstracts, and extracted text for all NIPS papers during 1987-2016. In particular the file `paper_authors.csv` contains a sparse matrix of paper coauthors.

### 6.2 NIPS words (1987-2015)

The following website:

`https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015`

collects the distribution of words in the full text of the NIPS conference papers published from 1987 to 2015. The dataset is in the form of a 11463 x 5812 matrix of word counts, containing 11463 words and 5811 NIPS conference papers (the first column contains the list of words). Each column contains the number of times each word appears in the corresponding document. The names of the columns give information about each document and its timestamp in the following format: `Xyear_paperID`.

## 7 Jiashun Jin's data on Coauthorship and Citation Networks for Statisticians

Thanks to Prof. Jiashun Jin at CMU, who provides his collection of citation and coauthor data for statisticians. The data set covers all papers between 2003 and the first quarter of 2012 from the Annals of Statistics, Journal of the American Statistical Association, Biometrika and Journal of the Royal Statistical Society Series B. The paper corrections and errata are not included. There are 3607 authors and 3248 papers in total. The zipped data file (14M) can be found at

`http://math.stanford.edu/~yuany/course/data/jiashun/Jiashun.zip`

with an explanation file

`http://math.stanford.edu/~yuany/course/data/jiashun/ReadMe.txt`

With the aid of Mr. LI, Xiao, a subset consisting 35 COPSS award winners (`https://en.wikipedia.org/wiki/COPSS_Presidents%27_Award`) up to 2015, is contained in the following file

`http://math.stanford.edu/~yuany/course/data/copss.txt`

An example was given in the following article, A Tutorial of Libra: R Package of Linearized Bregman Algorithms in High Dimensional Statistics, downloaded at

`http://math.stanford.edu/~yuany/course/reference/Libra_Tutorial_springer.pdf`

The citation of this dataset is: *P. Ji and J. Jin. Coauthorship and citation networks for statisticians. Ann. Appl. Stat. Volume 10, Number 4 (2016), 1779-1812*, (`http://projecteuclid.org/current/euclid.aoas`)

## 8 Co-appearance data in novels: Dream of Red Mansion and Journey to the West

A 374-by-475 binary matrix of character-event can be found at the course website, in .XLS, .CSV, .RData, and .MAT formats. For example the RData format is found at

`http://math.stanford.edu/~yuany/course/data/dream.RData`

with a readme file:

`http://math.stanford.edu/~yuany/course/data/dream.Rd`

as well as the .txt file which is readable by R command `read.table()`,

`http://math.stanford.edu/~yuany/course/data/HongLouMeng374.txt`

`http://math.stanford.edu/~yuany/course/data/readme.m`

Thanks to Ms. WAN, Mengting, who helps clean the data and kindly shares her BS thesis for your reference

`http://math.stanford.edu/~yuany/report/WANMengTing2013_HLM.pdf`

Moreover you may find a similar matrix of 302-by-408 for the Journey to the West (by Chen-En Wu) at:

`http://math.stanford.edu/~yuany/course/data/west.RData`

whose matlab format is saved at

`http://math.stanford.edu/~yuany/course/data/xiyouji.mat`

## 9 SNPs Data

This dataset contains a data matrix $X \in \mathbb{R}^{p\times n}$ 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://math.stanford.edu/~yuany/course/ceph_hgdp_minor_code_XNA.txt.zip`

which is big (151MB in zip and 2GB original txt). 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://math.stanford.edu/~yuany/course/data/HGDP_region.mat`

More detailed information about these persons in the dataset can be also found at

`http://math.stanford.edu/~yuany/course/data/HGDPid_populations_ALL.xls`

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

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

## 10 Protein Folding

Consider the 3D structure reconstruction based on incomplete MDS with uncertainty. Data file:

`http://math.stanford.edu/~yuany/course/data/protein3D.zip`

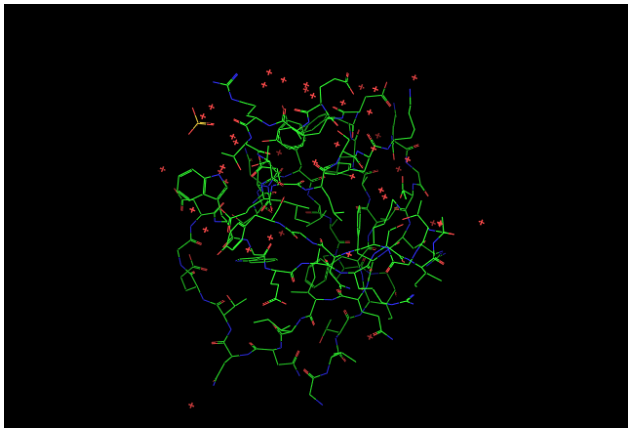

Figure 1: 3D graphs of file PF00018_2HDA.pdf (YES_HUMAN/97-144, PDB 2HDA)

In the file, you will find 3D coordinates for the following three protein families:

PF00013 (PCBP1_HUMAN/281-343, PDB 1WVN),

PF00018 (YES_HUMAN/97-144, PDB 2HDA), and

PF00254 (O45418_CAEEL/24-118, PDB 1R9H).

For example, the file `PF00018_2HDA.pdb` contains the 3D coordinates of alpha-carbons for a particular amino acid sequence in the family, YES_HUMAN/97-144, read as

`VALYDYEARTTEDLSFKKGERFQIINNTEGDWWEARSIATGKNGYIPS`

where the first line in the file is

97 V 0.967 18.470 4.342

Here

- '97': start position 97 in the sequence
- 'V': first character in the sequence

- $[x, y, z]$: 3D coordinates in unit $\mathring{A}$.

Figure 1 gives a 3D representation of its structure.

Given the 3D coordinates of the amino acids in the sequence, one can computer pairwise distance between amino acids, $[d_{ij}]^{l\times l}$ where $l$ is the sequence length. A *contact map* is defined to be a graph $G_\theta = (V, E)$ consisting $l$ vertices for amino acids such that and edge $(i, j) \in E$ if $d_{ij} \leq \theta$, where the threshold is typically $\theta = 5\mathring{A}$ or $8\mathring{A}$ here.

Can you recover the 3D structure of such proteins, up to an Euclidean transformation (rotation and translation), given noisy pairwise distances restricted on the contact map graph $G_\theta$, i.e. given noisy pairwise distances between vertex pairs whose true distances are no more than $\theta$? Design a noise model (e.g. Gaussian or uniformly bounded) for your experiments.

When $\theta = \infty$ without noise, classical MDS will work; but for a finite $\theta$ with noisy measurements, SDP approach can be useful. You may try the matlab package SNLSDP by Kim-Chuan Toh, Pratik Biswas, and Yinyu Ye, downladable at `http://www.math.nus.edu.sg/~mattohkc/SNLSDP.html`.