

Mini-Project 1

*Instructor: Yuan Yao**Due: 11:59pm Monday 10 Apr, 2021*

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 FOUR 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>

and a sample file at

<https://arxiv.org/pdf/1606.04930.pdf>

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. Reproducible source codes may be submitted via a <https://github.com/> link, through email as a zip file, or as an appendix of the report 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.

<https://yao-lab.github.io/data/snp452-data.mat>

or in R:

<https://yao-lab.github.io/data/snp500.Rda>

4 US Crime Data

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

<https://yao-lab.github.io/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

5 NIPS paper datasets

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

5.1 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.

5.2 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.

6 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

<https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/dream.RData>

with a readme file:

<https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/dream.Rd>

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

<https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/HongLouMeng374.txt>

<https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/README.md>

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

https://yao-lab.github.io/reference/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:

<https://github.com/yuany-pku/journey-to-the-west>

with R data format:

<https://github.com/yuany-pku/journey-to-the-west/blob/master/west.RData>

and Excel format:

<https://github.com/yuany-pku/journey-to-the-west/blob/master/xiyouji.xls>

7 SNPs Data

This dataset contains a data matrix $X \in \mathbb{R}^{n \times p}$ of about $p = 650,000$ columns of SNPs (Single Nucleid Polymorphisms) and $n = 1064$ rows of peoples around the world (but there are 21 rows mostly with missing values). Each element is of three choices, 0 (for 'AA'), 1 (for 'AC'), 2 (for 'CC'), and some missing values marked by 9.

<https://drive.google.com/file/d/1KMLPEG91mnzdK2pU1q2Bkj0x2BsaZy9s/view?usp=sharing>

which is big (151MB in zip and 2GB original txt). A fast access in the mainland China can be downloaded from:

https://pan.baidu.com/s/1jrv_UfbwWpi_-x5Rg1XS1A

with password 678e. Moreover, the following file contains the region where each people comes from, as well as two variables `ind1` and `ind2` such that $X(\text{ind1}, \text{ind2})$ removes all missing values.

https://github.com/yao-lab/yao-lab.github.io/blob/master/data/HGDP_region.mat

Another cleaned dataset is due to Quanhua MU and Yoonhee Nam:

- Genotyped data of the 1043 (n) subjects. 0(AA), 1(AC), 2(CC). Missing values are removed, only autosomal SNPs were selected ($p \approx 400K$). Google drive link:
https://drive.google.com/file/d/1a9I8_akfCMHBRrPMdnWkjyL9fKcQbJJq/view?usp=sharing
or <https://pan.baidu.com/s/1vDi0cLWl6GiWgm7icaZy-w> with password b5mv.
- Sample Information of 1043 subjects. Google drive link:
https://drive.google.com/file/d/11Q-8B57WDQnrIV92b-h_WLqDGviiYsm2/view?usp=sharing

A good reference for this data can be the following paper in Science,

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

Explore the genetic variation of those persons with their geographic variations, by MDS/PCA. Since p is big, explore random projections for dimensionality reduction.

8 Robust PCA: Video Clip

The following video clip (shoppingmall) has been widely used in literature for rank-sparsity decomposition of matrices. You may download the Matlab .mat file (50MB) from the following:

<https://drive.google.com/file/d/1CuVAG3uWnwq6QmI3vARUiz0F01Ubfz9k/view?usp=sharing>

The original .avi file (234MB) can be downloaded at

https://drive.google.com/file/d/10-wwU110fzzgvVF_YX0E1bEuU2Q9hGNG/view?usp=sharing

For those students in mainland China, another fast access of the data can be found at

<https://pan.baidu.com/s/1CNSBhueMLpLiD7gxVpQs0A>
with access password z9f6.

9 Protein Folding

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

<http://yao-lab.github.io/data/protein3D.zip>

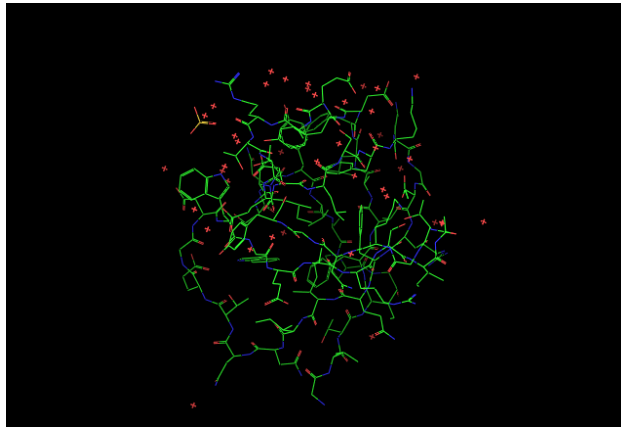


Figure 1: 3D graphs of file PF00018_2HDA.pdb (YES_HUMAN/97-144, PDB 2HDA)

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

```
VALYDYEARTTEDLSFKKGERFQIINNTEGDWWEARSIIATGKNGYIPS
```

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 \AA .

Figure 1 gives a 3D representation of its structure.

Given the 3D coordinates of the amino acids in the sequence, one can compute 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 of l vertices for amino acids such that an edge $(i, j) \in E$ if $d_{ij} \leq \theta$, where the threshold is typically $\theta = 5\text{\AA}$ or 8\AA 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_θ , i.e. given noisy pairwise distances between vertex pairs whose true distances are no more than θ ? 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 θ with noisy measurements, SDP approach can be useful. You may try the matlab package SNLSDP by Kim-Chuan Toh, Pratik Biswas, and Yinyu Ye, or the facial reduction speed up by Nathan Krislock and Henry Wolkowicz. Just for your reference, the following version SNLSDP is collected and updated by Mengyue ZHA in the class,

<https://github.com/Dolores2333/MATH5473/tree/main/HW5/SNLSDP>

For python users, you may try the Python version of CVX (CVXPY): <https://www.cvxpy.org/install/index.html>.