Topological and Geometric Data Reduction **16 April, 2025**

Final Project

Instructor: Yuan Yao *Due: 23:59 April 29, 2025*

1 Project Requirement and Datasets

In the below, we list some candidate datasets for your reference. You are also encouraged to work on your own datasets in the final project, upon the approval of the instructor.

1. Pick up ONE (or more if you like) favourite 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 must submit:

   (a) *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`

   *(b)[1] *ONE short presentation video within 10 mins*, e.g. in Youtube or Bilibili link. You may submit your presentation slides together with the vieo link to help understanding.

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. Upon the start, submit your grouping information (representer/member) to the following address (datascience.hw@gmail.com) with Title: CSIC 5011: Project 2 [grouping]. Then we shall create public folders and you should submit your report to your group folder under subfolder `[GROUP]/report`.

[1] We plan to do final project presentation in the last lecture.

2 Some classical image datasets

2.1 Face dataset

The following dataset contains 33 faces of the same person ($Y \in \mathbb{R}^{112\times 92\times 33}$) in different angles,

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

You may create a data matrix $X \in \mathbb{R}^{n\times p}$ where $n = 33, p = 112 \times 92 = 10304$ (e.g. `X=reshape(Y,[10304,33])';` in matlab). You may try Python package

`https://scikit-learn.org/stable/modules/manifold.html#`

For example,

1. Explore the Diffusion map, or the second smallest eigenvector of Markov Chains defined on the point cloud data, to order the faces, i.e., let $W_{ij} = \exp(-\|x_i - x_j\|^2/t)$ with $D = \mathrm{diag}(\sum_j W_{ij})$ and define $L = D^{-1}W - I$, clearly $\lambda_0 = 0$ and take the (second) smallest nonzero eigenvalue λ_1 with corresponding eigenvector v_1, sort the faces by values $v_1(i)$, $i = 1, \ldots, n$.
2. Explore the MDS-embedding of the 33 faces on top two eigenvectors: order the faces according to the top 1st eigenvector and visualize your results with figures.
3. Explore the ISOMAP-embedding of the 33 faces on the $k = 5$ nearest neighbor graph and compare it against the MDS results. Note: you may try Tenenbaum's Matlab code `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/isomapII.m`
4. Explore the LLE/MLLE-embedding of the 33 faces on the $k = 5$ nearest neighbor graph and compare it against ISOMAP. Note: you may try the following Matlab/Python code `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/lle.m`,
5. Explore the LSTA-embedding of the 33 faces on the $k = 5$ nearest neighbor graph and compare it against ISOMAP. Note: you may try the following Python code `https://scikit-learn.org/stable/modules/manifold.html#local-tangent-space-alignment`
6. Explore the 2-D t-SNE embedding of the 33 faces.

You might explore larger datasets with other manifold learning methods below.

2.2 PubFig dataset

The PubFig dataset is at

`http://www.cs.columbia.edu/CAVE/databases/pubfig/`

2.3 MNIST dataset

Yann LeCun's website contains original MNIST dataset of 60,000 training images and 10,000 test images.

`http://yann.lecun.com/exdb/mnist/`

There are various ways to download and parse MNIST files. For example, Python users may refer to the following website:

`https://github.com/datapythonista/mnist`

or MXNET tutorial on mnist

`https://mxnet.incubator.apache.org/tutorials/python/mnist.html`

2.4 Fashion-MNIST dataset

Zalando's Fashion-MNIST dataset of 60,000 training images and 10,000 test images, of size 28-by-28 in grayscale.

`https://github.com/zalandoresearch/fashion-mnist`

2.5 Hand-written Digits

The website

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

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

2.6 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`

2.7 Animal Sleeping Data

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

`https://yao-lab.github.io/data/sleep1.csv`

2.8 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`

3 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/1KMLPEG91mnzdK2pUlq2BkjOx2BsaZy9s/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 $X($`ind1`$,$ `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.

4 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/1CuVAG3uWnwq6QmI3vARUizOF01Ubfz9k/view?usp=sharing`

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

`https://drive.google.com/file/d/10-wwUl10fzzgvVF_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/1CNSBhueMLpLiD7gxVpQsOA`

with access password `z9f6`.

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 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

`https://yao-lab.github.io/data/jiashun/Jiashun.zip`

with an explanation file

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

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

`https://arxiv.org/abs/1604.05910`

with the associated R package Libra:

`https://cran.r-project.org/web/packages/Libra/index.html`

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`)

7 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`

8 Protein Folding

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

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

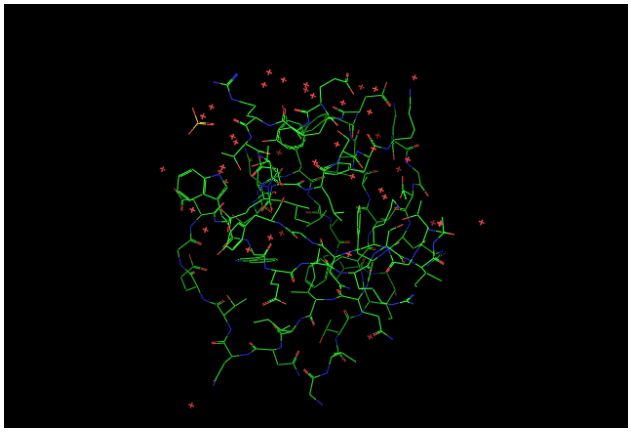

Figure 1: 3D graphs of file PF00018_2HDA.pdf (YES_HUMAN/97-144, PDB 2HDA)

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_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`.

9 Human Prefrontal Cortex Development Data

This dataset contains a single cell gene expression matrix $X \in \mathbb{R}^{n\times p}$ of $n = 24153$ genes and $p = 2394$ cells. Each value is in the unit of transcript-per-million (TPM). The file of the dataset is in tab-delimited text format, where each row represents one gene (the first row is the cell ID) and

each column represents one cell (the first column is the gene name). The size is about 33.2 Mb in gzipped format and about 160 Mb after decompression. Link to download the data:

`https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE104276&format=file&file=GSE104276%5Fall%5Fpfc%5F2394%5FUMI%5FTPM%5FNOERCC%2Exls%2Egz`

The single cells were collected in the human embryonic prefrontal cortex (PFC) from gestational weeks (GW) 8 to 26. More specifically, the cells were collected at GW8, GW9, GW10, GW12, GW13, GW16, GW19, GW23 and GW26, and included three and two biological replicates at GW10 and GW23, respectively. In the gene expression matrix, the cell ID contains both the time point and the donor information. For example, cell "GW10_PFC1_sc35" indicates that the cell was collected at gestational week 10 from donor 1. To learn more about the dataset you can visit the Gene Expression Omnibus (GSE104276) or read the original publication in *Nature*:

- Zhong, S., Zhang, S., Fan, X. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex, *Nature* **555**, 524-528 (2018). https://doi.org/10.1038/nature25980.

Can you identify subgroups of cells from the data, and trace the developmental trajectories of these cells? Dimensionality reduction methods and single cell topological data analysis methods may help.

10 Dynamic Simplicial Complexes of Teacher-Student Contacts

Infection diseases like COVID-19 now influenced everyone's life and are a challenge for human to fight. The following data set contains the temporal network of contacts between the children and teachers in a primary school used in the study published in BMC Infectious Diseases 2014, 14:695.

`http://www.sociopatterns.org/datasets/primary-school-temporal-network-data/`

The file contains a tab-separated list representing the active contacts during 20-second intervals of the data collection. Each line has the form "`t i j Ci Cj`", where `i` and `j` are the anonymous IDs of the persons in contact, `Ci` and `Cj` are their classes, and the interval during which this contact was active is $[\mathtt{t} - 20s, \mathtt{t}]$. If multiple contacts are active in a given interval, you will see multiple lines starting with the same value of `t`. Time is measured in seconds.

Based on such data, simplices are extracted through cliques of simultaneous contacts in the following directory by Austin Benson et al.:

`https://github.com/arbenson/ScHoLP-Data/tree/master/contact-primary-school`

Specifically, for every unique timestamp in the dataset, a simplex is constructed for every maximal clique amongst the contact edges that exist for that timestamp. Timestamps were recorded in 20 second intervals.

You may construct simplicial complexes (a collection of simplices that is closed under inclusion) from such data and study its topological properties (e.g. Betti numbers) with dynamic changes.

For those who are studying genomics can explore the phylogenetic trees and networks for genomics data of hCoV-19 at:

`https://www.gisaid.org/`

11 PageRank and Chinese Universities

The following dataset contains Chinese (mainland) University Weblink during 12/2001-1/2002,

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

where `rank_cn` is the research ranking of universities in that year, `univ_cn` contains the webpages of universities, and `W_cn` is the link matrix whose $(i, j) - th$ element gives the number of links from university i to j.

1. Compute PageRank with Google's hyperparameter $\alpha = 0.85$;
2. Compute HITS authority and hub ranking;
3. Compare these rankings against the research ranking (you may consider Spearman's ρ and Kendall's τ to compare different rankings);
4. Compute extended PageRank with various hyperparameters $\alpha \in (0, 1)$, investigate its effect on ranking.

For your reference, an implementation of PageRank and HITs can be found at

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/pagerank.m`

The following academic website link collects more countries with university links, for further explorations:

`http://cybermetrics.wlv.ac.uk/database/`

12 Crowdsourced Ranking Data on Allourideas

The following datasets are crowdsourced pairwise ranking from platform Allourideas by Professor Mathew Salganik of Princeton Sociology. You may explore it with HodgeRank etc.

12.1 World College Rankings

The following website hosts the crowdsourcing task on pairwise ranking on 270 universities in the world:

`http://www.allourideas.org/worldcollege`

Up to Nov 26, 2017, the following dataset is collected at github:

`https://github.com/yuany-pku/data/tree/master/allourideas/allourideas_worldcollege`

where you may find

- explanation of data file formats: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/allourideas%20-%20download%20your%20data.pdf`
- 270 universities: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/wikisurvey_colleges_candidates_2017-11-26T07_14_53Z.csv`
- all valid votings: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/wikisurvey_colleges_votes_2017-11-26T07_15_02Z.csv`
- all nonvotings: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/wikisurvey_colleges_nonvotes_2017-11-26T07_15_30Z.csv`

This dataset has been used for various studies, e.g. Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. False Discovery Rate Control and Statistical Quality Assessment of Annotators in Crowdsourced Ranking, ICML 2016, in `https://arxiv.org/abs/1605.05860v1`. An old dataset cleaned by Prof. Qianqian Xu from CAS can be found at

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/college.csv`

12.2 Human Age Ranking

The following dataset is kindly provided by Qianqian Xu, CAS, for the exploration on class.

The dataset is contained in the following zip file.

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/age.zip`

where you may find

1. `readme.txt`: description of data
2. `Agedata.mat`: data file collected
3. `Groundtruth.mat`: Groundtruth
4. `30 images.zip`: 30 human face images of different ages

The basic problem is to rank the faces according to the ages, using all the information collected so far. A simple sub-problem is rank aggregation of ages from pairwise comparisons. If you are interested, you can try some generalized linear models (Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. HodgeRank on Random Graphs for Subjective Video Quality Assessment. IEEE Transactions on Multimedia, 14(3):844-857, 2012, `https://github.`

com/yao-lab/yao-lab.github.io/blob/master/reference/TMM12-final.pdf) on this dataset, such as uniform model, Bradley-Terry model, Thurstone-Mosteller model, and Angular transform model. Compare maximum likelihood estimators and least square ones. The source code of this paper can be found at

`https://github.com/qianqianxu010/TMM2012`

A recent study with wider data is: Qianqian Xu, Jiechao Xiong, Xiaochun Cao, Qingming Huang, Yuan Yao, From Social to Individuals: a Parsimonious Path of Multi-level Models for Crowdsourced Preference Aggregation, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 41(4):844-856, 2019, where the source codes can be downloaded at

`https://github.com/qianqianxu010/TPAMI2018`