

Mini-Project 1.

*Instructor: Yuan Yao**Due: Tuesday May 17, 2014*

1 Mini-Project Requirement and Datasets

This project aims to exercise the tools in the class, such as random projections, robust PCA, sparse PCA, 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 *poster* report, *with a clear remark on each person's contribution*. A sample poster file with PKU logo can be found at http://math.stanford.edu/~yuany/course/reference/poster_v5.pdf whose source LATEX codes can be downloaded at <http://math.stanford.edu/~yuany/course/reference/pkuposter.zip>
3. In the report, (1) design or raise your scientific problems (a good problem is sometimes more important than solving it); (2) show your results with your careful analysis of the results toward answering your problem. 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 Teaching Assistant (TA), Huizhuo Yuan (datascience_hw@126.com).

2 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>

3 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. arXiv:1410.2840, 2014.* As the paper has not been formally published yet, please do not use the dataset outside this class or for any kinds of publications without the permission of the authors.

4 Drug Efficacy Data

Thanks to Prof. Xianting Ding at Shanghai Jiao Tong University and Prof. Chih-Ming Ho from University of California at Los Angeles, we have the following datasets on combinatorial drug efficacy.

The first dataset consists of two experiments, all with the same 4 drugs in cell lines for attacking leukemia, with 256 experiments of combinatorial drug dosage at 4 levels. The response is the therapeutic window measuring the efficacy with a trade-off by toxicity.

http://math.stanford.edu/~yuany/course/data/Ding_4drugs.xlsx

whose drugs are explained in

http://math.stanford.edu/~yuany/course/data/Ding_4drugs_readme.pdf

Can you find a good prediction of drug response efficacy using those combinatorial dosage levels? It was suggested that quadratic polynomials at logarithmic dosage levels are good models in personalized medicine, e.g. the following cover paper in *Science Translation Medicine*:

<http://stm.sciencemag.org/content/8/333/333ra49>

with a sample 14 drug efficacy at level 2 experiment data in liver transplant:

<http://math.stanford.edu/yuany/course/data/TB-FSC-03A-data.xlsx>

5 Heart PCI Operation Effect Prediction

The following data, provided by Dr. Jinwen Wang at Anzhen Hospital,

http://math.stanford.edu/~yuany/course/data/heartData_20140401.xlsx

contains 2581 patients with 73 measurements (inputs) as well as a response variable indicating if after the heart operation there is a null-reflux state. This is a classification problem, with a challenge from the large amount of missing values. Sheet 3 and 4 in the file contains some explanation of the data and variables.

The problems are listed here:

1. The inputs (covariates) are of three kinds, measurements upon check-in, measurements before PCI operation, and measurements in PCI operations. For doctors, it is desired to find a prediction model based on measurements before the operation (including check-in). Sheet 2 in the file contains only such measurements.

The following two reports by LV, Yuan and LI, Xiao, respectively, might be interesting to you:

<http://math.stanford.edu/~yuany/course/reference/MSThesis.LvYuan.pdf>

<http://arxiv.org/abs/1511.04656>

2. It is also an interesting problem how to predict the effect based on all measurements, with lots of missing values. Sheet 1 contains the full measurements. There are some good work by previous students, which are listed here for your reference:

The following report by MIAO, Wang and LI, Yanfang, pioneers in missing value treatment.

http://math.stanford.edu/~yuany/course/reference/MiaoLi2013S_project01.pdf

6 Identification of Raphael's paintings from the forgeries

The following data, provided by Prof. Yang WANG from HKUST,

<https://drive.google.com/folderview?id=0B-yDtwSjhaSCZ2FqN3AxQ3NJNTA&usp=sharing>

contains a 28 digital paintings of Raphael or forgeries. Note that there are both jpeg and tiff files, so be careful with the bit depth in digitization. The following file

<https://docs.google.com/document/d/1tMaaSIrYwNFZZ2cEJdx1DfFscIfERd5Dp2U7K1ekjTI/edit>

contains the labels of such paintings, which are

- 1 Maybe Raphael - Disputed
- 2 Raphael
- 3 Raphael
- 4 Raphael
- 5 Raphael
- 6 Raphael
- 7 Maybe Raphael - Disputed
- 8 Raphael
- 9 Raphael
- 10 Maybe Raphael - Disputed
- 11 Not Raphael
- 12 Not Raphael
- 13 Not Raphael
- 14 Not Raphael
- 15 Not Raphael

- 16 Not Raphael
- 17 Not Raphael
- 18 Not Raphael
- 19 Not Raphael
- 20 My Drawing (Raphael?)
- 21 Raphael
- 22 Raphael
- 23 Maybe Raphael - Disputed
- 24 Raphael
- 25 Maybe Raphael - Disputed
- 26 Maybe Raphael - Disputed
- 27 Raphael
- 28 Raphael

Can you exploit the known Raphael vs. Not Raphael data to predict the identity of those 6 disputed paintings (maybe Raphael)? The following student poster report seems a good exploration

http://math.stanford.edu/~yuany/course/2015.fall/poster/Raphael_LI%2CYue_1300010601.pdf

The following paper by Haixia Liu, Raymond Chan, and me studies Van Gogh's paintings which might be a reference for you:

<http://dx.doi.org/10.1016/j.acha.2015.11.005>

7 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://www.math.pku.edu.cn/teachers/yaoy/data/snp452-data.mat>

8 Hand-written Digits

The website

<http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.digits/>
contains images of 10 handwritten digits ('0', ..., '9');

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(ind1,ind2)` removes all missing values.

`http://math.stanford.edu/~yuany/course/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`

You can login my server:

`ssh einstein@162.105.205.92`

using the password I provided on class. On the read only folder `/data/snp/`, you will find all the data in both `.txt` and `.mat` (`data.mat`, `HGDP_region.mat`, `readme.m`).