**A Mathematical Introduction to Data Analysis** **27 Feb, 2017**

# Mini-Project 1

*Instructor: Yuan Yao* *Due: Monday 13 Mar, 2017*

## 1 Mini-Project Requirement and Datasets

This project aims to exercise the tools in the class, such as PCA, 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 *poster* report, *with a clear remark on each person's contribution.*
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: Math 6380: Proj 1.

## 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 NIPS paper datasets

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

## 4.1 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.

## 4.2 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`.

# 5 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`

## 6 Drug Sensitivity Data by Cleave

The following dataset is kindly provided by Cleave Co. Ltd. USA, for the exploration on class. **Please keep its use only in this class and any publication will be subject to the approval of Cleave.**

The dataset is contained in the following zip file (73M).

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

where you may find

1. `data explanation.pptx`: description of data in pptx
2. `data for Yuan Yao.xlsx`: data file
3. `Gene set collection 1 for Yuan Yao.txt`: gene set collection
4. `Gene set collection 2 for Yuan Yao.txt`: gene set collection
5. `reference`: a folder contains a survey paper on 40+ machine learning algorithms as well as some source codes – *Nature Biotechnology 32, 1202–1212 (2014)* (`http://www.nature.com/nbt/journal/v32/n12/full/nbt.2877.html`)

The basic problem is to predict the drug response `IC50 within 72 hours`, using all the information collected so far, introduced by Ms. Lijing Wang with slides

`http://math.stanford.edu/~yuany/course/2016.spring/cleave_lijing.pdf`

as well as our CPH'2017 poster

`http://math.stanford.edu/~yuany/publications/poster_CleaveBioCPH2017_ForReview.pdf`

where the crucial discovery is that recursive variable selection by LASSO is more effective than one-stage LASSO.

## 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');