A Mathematical Introduction to Data Science **November 4, 2014**

Mini-Project 1.

Instructor: Yuan Yao *Due: Tuesday November 18, 2014*

The problem below marked by * is optional with bonus credits.

1. *Manifold Learning*: The following codes by Todd Wittman contain major manifold learning algorithms talked on class.

   `http://www.math.pku.edu.cn/teachers/yaoy/Spring2011/matlab/mani.m`

   Precisely, eight algorithms are implemented in the codes: MDS, PCA, ISOMAP, LLE, Hessian Eigenmap, Laplacian Eigenmap, Diffusion Map, and LTSA. The following nine examples are given to compare these methods,

   (a) Swiss roll;
   (b) Swiss hole;
   (c) Corner Planes;
   (d) Punctured Sphere;
   (e) Twin Peaks;
   (f) 3D Clusters;
   (g) Toroidal Helix;
   (h) Gaussian;
   (i) Occluded Disks.

   Run the codes for each of the nine examples, and analyze the phenomena you observed.

2. *RPCA*: Construct a random rank-r matrix: let $A \in \mathbb{R}^{m\times n}$ with $a_{ij} \sim \mathcal{N}(0,1)$ whose top-r singular value/vector is λ_i, $u_i \in \mathbb{R}^m$ and $v_i \in \mathbb{R}^n$ $(i = 1, \ldots, r)$, define $L = \sum_{i=1}^r u_i v_i^T$. Construct a sparse matrix E with p percentage ($p \in [0,1]$) nonzero entries distributed uniformly. Then define

$$M = L + E.$$

   (a) Set $m = n = 20$, $r = 1$, and $p = 0.1$, use Matlab toolbox CVX to formulate a semi-definite program for Robust PCA of M:

$$\begin{aligned} \min \quad & \frac{1}{2}(\mathrm{trace}(W_1) + \mathrm{trace}(W_2)) + \lambda\|S\|_1 \\ s.t. \quad & L_{ij} + S_{ij} = X_{ij}, \quad (i,j) \in E \\ & \begin{bmatrix} W_1 & L \\ L^T & W_2 \end{bmatrix} \succeq 0, \end{aligned} \tag{1}$$

   where you can use the matlab implementation in lecture notes as a reference;

(b) Choose different parameters $p \in [0, 1]$ to explore the probability of successful recover;

(c) Increase r to explore the probability of successful recover;

(d) * Increase m and n to values beyond 50 will make CVX difficult to solve. In this case, use the Augmented Lagrange Multiplier method, e.g. in E. J. Candes, X. Li, Y. Ma, and J. Wright (2009) "Robust Principal Component Analysis?". Journal of ACM, 58(1), 1-37 (`http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/rpca.pdf`). Make a code yourself (just a few lines of Matlab or R) and test it for $m = n = 1000$. A convergence criterion often used can be $\|M - \hat{L} - \hat{S}\|_F / \|M\|_F \leq \epsilon$ ($\epsilon = 10^{-6}$ for example).

3. *SPCA*: Define three hidden factors:

$$V_1 \sim \mathcal{N}(0, 290), \quad V_2 \sim \mathcal{N}(0, 300), \quad V_3 = -0.3V_1 + 0.925V_2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1),$$

where V_1, V_2, and ϵ are independent. Construct 10 observed variables as follows

$$X_i = V_j + \epsilon_i^j, \quad \epsilon_i^j \sim \mathcal{N}(0, 1),$$

with $j = 1$ for $i = 1, \ldots, 4$, $j = 2$ for $i = 5, \ldots, 8$, and $j = 3$ for $i = 9, 10$ and ϵ_i^j independent for $j = 1, 2, 3$, $i = 1, \ldots, 10$.

The first two principal components should be concentrated on (X_1, X_2, X_3, X_4) and (X_5, X_6, X_7, X_8), respectively. This is an example given by H. Zou, T. Hastie, and R. Tibshirani, Sparse principal component analysis, J. Comput. Graphical Statist., 15 (2006), pp. 265-286.

(a) Compute the true covariance matrix Σ (and the sample covariance matrix with n examples, say $n = 1000$);

(b) Compute the top 4 principal components of Σ using eigenvector decomposition (by Matlab or R);

(c) Use Matlab CVX toolbox to compute the first *sparse* principal component by solving the SDP problem

$$\begin{array}{ll} \max & \text{trace}(\Sigma X) - \lambda \|X\|_1 \\ s.t. & \text{trace}(X) = 1 \\ & X \succeq 0 \end{array}$$

Choose $\lambda = 0$ and other positive numbers to compare your results with normal PCA;

(d) Remove the first sparse PCA from Σ and compute the second sparse PCA with the same code;

(e) Again compute the 3rd and the 4th sparse PCA of Σ and compare them against the normal PCAs.

(f) * Construct an example with 200 observed variables which is hard to deal with by CVX. In this case, use the Augmented Lagrange Multiplier method by Allen Yang et al. (UC Berkeley) whose Matlab codes can be found at `http://www.eecs.berkeley.edu/~yang/software/SPCA/SPCA_ALM.zip`.

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 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 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), Jiechao Xiong (datascience_hw@126.com).

1.1 The Characters in A Dream of Red Mansion

A 376-by-475 matrix of character-event can be found at the course website, in .XLS, .CSV, and .MAT formats. For example the Matlab format is found at

`http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/hongloumeng376.mat`

with a readme file:

`http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/readme.m`

Thanks to Ms. WAN, Mengting (now at UIUC), an update of data matrix consisting 374 characters (two of 376 are repeated) which is readable by R read.table() can be found at

`http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/HongLouMeng374.txt`

She also kindly shares her BS thesis for your reference

`http://www.math.pku.edu.cn/teachers/yaoy/reference/WANMengTing2013_HLM.pdf`

1.2 A Journal to the West

On course website, you may also find the link to this dataset with a 302-by-408 matrix, whose matlab format is saved at

`http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/xiyouji/xiyouji.mat`

For your reference, here is a project presentation by Mr. LI, Liying (at PKU) which gives an analysis based on PCA

`http://www.math.pku.edu.cn/teachers/yaoy/reference/LiyingLI_Xiyouji2012_slides.pdf`

1.3 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://www.math.pku.edu.cn/teachers/yaoy/data/snp452-data.mat`

1.4 Hand-written Digits

The website

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

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

1.5 Air Quality Weibo Data

(courtesy of Prof. Xiaojin Zhu from University of Wisconsin at Madison) You can login my server:

`ssh einstein@162.105.205.92`

using the password I provided on class.

On the read-only folder `/data/AQweibo/`, the `AQICityData/` directory contains the Weibo posts, the AQI for 108 cities with (AQI) information during the study period from 2013-11-18 to 2013-12-18 (both inclusive); Information for the spatiotempral bin (city,date) is in the directory `city_date/`. See `README.txt` for more information.

1.6 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://www.math.pku.edu.cn/teachers/yaoy/data/ceph_hgdp_minor_code_XNA.txt.zip`

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://www.math.pku.edu.cn/teachers/yaoy/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`

Attention: this last dataset is relatively big with about 2GB size.

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`).

1.7 Bird Flu Dataset

(courtesy of Steve Smale and Cissy) This dataset 162 H5N1 (bird flu) virus sequences discovered around the world:

`http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_seq162.txt`

Locations of such virus discovered are reported with latitude and longitude coordinates on the globe:

`http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_latgrat.txt`

Pairwise geodesic distances between these 162 sites are constructed as

`http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_geodist.txt`

A kernel-induced l_2-distances between 162 virus sequences are given in

`http://www.math.pku.edu.cn/teachers/yaoy/data/birdflu_l2dist.txt`