Statistical Learning **September 28, 2013**

Project I: PCA

Instructor: Yuan Yao *Due: October 16, 2013*

1 Requirement

1. Pick up ONE (or more if you like) favorite problem below to attack. 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, show your results with your careful analysis of the results. 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 Assistants (TA), Junliang Huang (jlhwung@gmail.com) and Qing Wang (wangqing.linus@gmail.com).

2 Social Network Data

2.1 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`

Among various choices of analysis, with this data matrix X, you may form a weighted graph $W = X * X'$, pursue PCA of X.

2.2 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`

3 SNPs Data

The following dataset contains 650K-SNPs by 1000-persons scattered around the globe, published by Human Genome Diversity Project

`http://www.cephb.fr/en/hgdp/`

You may choose other SNPs datasets at your interests. PCA has been used to disclose geographic information (mapping the Human Evolution history) in literature

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

4 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`

You may use PCA to explore the 'invisible hands' of markets.