A Mathematical Introduction to Data Science **Sep. 15, 2017**

Homework 2. Random Projections

Instructor: Yuan Yao *Due: Open Date*

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

1. *SNPs of World-wide Populations:* 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.

   `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 $X(\texttt{ind1}, \texttt{ind2})$ removes all missing values.

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

   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.

2. *Phase Transition in Compressed Sensing:* Let $A \in \mathbb{R}^{n\times d}$ be a Gaussian random matrix, *i.e.* $A_{ij} \sim \mathcal{N}(0,1)$. In the following experiments, fix $d = 20$. For each $n = 1,\ldots,d$, and each $k = 1,\ldots,d$, repeat the following procedure 50 times:

   (a) Construct a sparse vector $x_0 \in \mathbb{R}^d$ with k nonzero entries. The locations of the nonzero entries are selected at random and each nonzero equals ± 1 with equal probability;

   (b) Draw a standard Gaussian random matrix $A \in \mathbb{R}^{n\times d}$, and set $b = Ax_0$;

   (c) Solve the following linear programming problem to obtain an optimal point $\hat{x}$,

   $$\begin{aligned} \min_x \quad & \|x\|_1 := \sum |x_i| \\ s.t. \quad & Ax = b, \end{aligned}$$

   for example, matlab toolbox `cvx` can be an easy solver;

   (d) Declare success if $\|\hat{x} - x_0\| \leq 10^{-3}$;

   After repeating 50 times, compute the success probability $p(n,k)$; draw a figure with x-axis for k and y-axis for n, to visualize the success probability. For example, matlab command `imagesc(p)` can be a choice.

   Can you try to give an analysis of the phenomenon observed? Tropp's paper mentioned on class may give you a good starting point.