# Statistics and Numerical Method — Problem Set #3 (due 11/11/2019)

### 1. Statistics with the Cauchy distribution (9 pts)

(1). Your programming language should have a standard random number generator that can generate a uniform distribution between $[0, 1]$. Utilizing this, write your own code to generate a Cauchy distribution with $\mu = 0$ and $\gamma = 1$. (1 pt)

(2). Calculate (analytically) the value of $q_{25}$ and $q_{75}$, as well as the median ($q_{50}$) of the distribution as a function of $\mu$ and $\gamma$. (1 pt)

(3). Using your generator, create 1000 numbers satisfying this Cauchy distribution. Sort your data, and find out the median and interquartile range from your data. Use this information to estimate $\mu$ and $\gamma$. (2 pts)

(4). Based on the 1000 data points, use bootstrap (say do it 1000 times) to estimate the 95% confidence interval of the quantities that you estimated in (2). (2 pts)

(5). Repeat the above but you have 10000 samples (instead of 1000). How much improvement do you get? (1 pts)

(6). Discuss the prospect of using the maximum likelihood method to infer the parameters $\mu$ and $\gamma$. (Write down the equations, but I am not requesting you to solve them.) (2 pts)

### 2. GRE physics test (7 pts)

Over a long time, students applying astronomy graduate programs in the US were required take the GRE physics (PGRE) test, which has been used as a criterion for student admission. A few years ago, a survey was conducted among people who were awarded prestigious postdoctoral fellowships in astronomy, assuming (not necessarily true) such fellowships represent one possible measure of "success" in academic career. The survey data on their PGRE scores are summarized in Table 1 expressed in terms of their percentile ranking. The survey was aimed at assessing whether the PGRE scores can be a good indicator of future career success. Obviously, there are many biasing factors that may make some survey results questionable, but for the purpose of our problem, we shall ignore them for simplicity.

(1). Visualize the data set in a histogram for each percentile bin ($0 - 10\%$, $10 - 20\%$, etc.). (1 pt)

(2). Design a test to answer: are the data consistent drawing randomly from a general population of students taking the PGRE test? Please give the p-value of your test. (2 pts)

(3). Design a test to answer: is there a difference in PGRE scores between male and female fellowship awardees? Please give the p-value of your test. (2 pts)

(4). If US graduate schools adopt a hard threshold and exclude all applicants whose PGRE scores are below 50%. Would you think this is a good idea? Why? (2 pts)

### 3. Non-negative matrix factorization (NMF) (16 pts)

Non-negative matrix factorization (NMF) was introduced by D.D. Lee and H.S. Seung (1999) in a *Nature* paper (https://www.nature.com/articles/44565) as a novel idea for matrix factorization.

Table 1: Statistics of GRE percentiles

| | $\mathbf{N_{total}}$ | <10% | <20% | <30% | <40% | <50% | <60% | <70% | <80% | <90% |
|---|---|---|---|---|---|---|---|---|---|---|
| All | 149 | 4 | 10 | 16 | 32 | 44 | 65 | 85 | 104 | 121 |
| Male | 75 | 1 | 3 | 6 | 13 | 15 | 25 | 34 | 47 | 59 |
| Female | 53 | 1 | 4 | 6 | 14 | 22 | 32 | 42 | 47 | 49 |
| Unspecified | 21 | 2 | 3 | 4 | 5 | 7 | 8 | 9 | 10 | 13 |

Different from other matrix factorization approaches, the NMF works for a non-negative matrix $V \in \mathbb{R}^{m\times n}$, and seeks for

$$V \approx WH \ , \tag{1}$$

where $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{r\times n}$, and both of them have all their matrix components non-negative. Usually $r \ll \min(m, n)$, so that $W$ is a "tall" matrix, and $H$ is a "fat" matrix. The interpretation is straightforward: individual columns of $V$ can be written as linear combinations of columns in $W$ with non-negative coefficients (which is given by $H$). The NMF has extensive applications in computer vision, audio signal processing, clustering analysis, recommender systems, etc., and is also becoming an emerging tool in astronomy (image processing).

In general, we do not expect the NMF to be exact. Rather, we are satisfied as long as $WH$ is sufficiently close to $V$. Therefore, the NMF problem is formulated as a constrained optimization problem as follows

$$\text{Minimize}: \ L(W, H) = \|V - WH\|_{\text{Fro}} \ , \quad \text{subject to } W, H > 0 \ . \tag{2}$$

There is also an alternative formulation (assuming Poisson-like noise), as used in Lee & Seung,

$$\text{Minimize}: Q(W, H) = \sum_{ij} (WH)_{ij} - V_{ij} \ln(WH)_{ij} \ , \quad \text{subject to } W, H > 0 \ . \tag{3}$$

We will see some applications of NMF later in this course (but it will not discuss the underlying algorithm). In this homework problem, you are asked to reproduce the result presented in Lee & Seung, using their original data. The data consist of 2419 gray images of human faces, and each image has a size of $19 \times 19$ pixels (so 361 elements). You can find the data file attached online, which is a plain txt file listing the pixel values one image after another. When you read the file, please display the images first (using a gray colormap) to make sure you see faces, which means you get the order of the arrays correctly (see Figure 1 as an example).[1]

(1). Write an NMF solver for this problem. Note that NMF should be conducted on a $361 \times 2419$ matrix so that each image is a column. The main parameter is $r$, namely, number of "features" (i.e. the dimension in $W$, $H$) retained in NMF, and the initial guesses $W^{(0)}$, $H^{(0)}$ should also be input parameters. Either of the two objective functions can be used for this problem, and for NMF algorithms, you can choose one of the following.

[1] You may also need to adjust the viewing angle so that the faces have the right orientation.

(a) Implement the original algorithm of Lee & Seung described in that paper that minimizes (3), or similar algorithms that minimize (2) (e.g., an update formula is provided in wikipedia).

(b) Properly formulate the problem and use any optimization packages available to you (e.g., the CVX family packages, NLOPT, etc.).

However, you are not allowed to directly call the NMF routine (e.g., in python/matlab/Julia) for this task. Please attach your code (and specify the optimization package in case you choose b), and also describe the stopping criterion used in your code (or in the package if you choose b). (6 pts)

(2). Choose $r = 10, 20, 30, 50$, and for each of these choices, start the iteration with each element of $W^{(0)}$, $H^{(0)}$ chosen from a uniform distribution in $[0, 1]$. Record the value of the objective function at each iteration. To illustrate your results, provide three main figures that show the following.

a). Among the 2419 faces, choose the 1st, 5th, 25th, 125th, 625th, and the 2419th faces. Make a figure that contains $6 \times 5$ images, with first column showing the original image, and the rest showing the reconstructed image (from $V' = WH$) for each value of $r$. (see Figure 1 for a possible way to arrange the images). (3 pts)

b). For $r = 30$, show another figure containing 30 images that correspond to the columns of the $W$ matrix obtained from your NMF. (2 pts) Briefly explain what you find. (1 pt)

c). For $r = 10, 20, 30, 50$, plot the value of the objective function as a function of iteration steps. (2 pts)

(3). Does the outcome of your NMF depend on initial guesses? Briefly explain why. (2 pts)

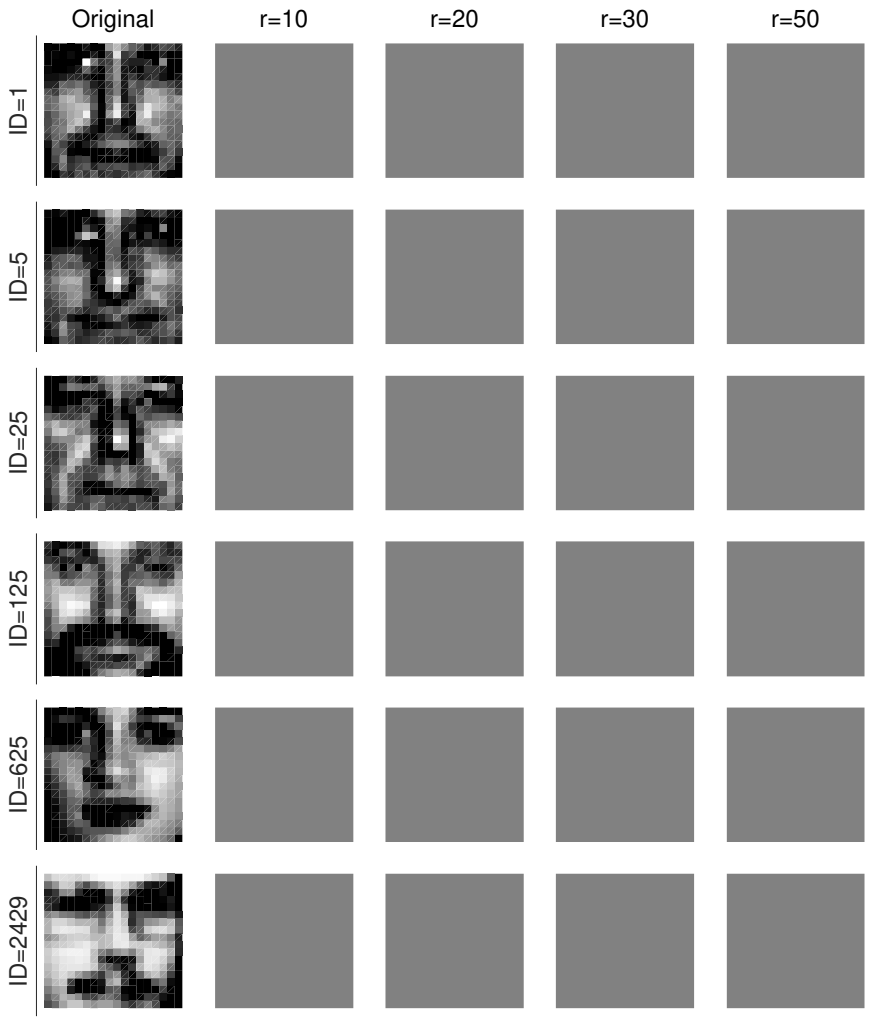


Figure 1: List of 6 sample faces together with results from NMF for problem 3-(2)-a).