Carnegie Mellon University
Department of Computer Science
15-826 Multimedia Databases and Data Mining
C. Faloutsos, Fall 2024

Homework 4 - Solutions
Due: <u>**pdf, on canvas, at 2:00pm, on 11/22/2024**</u>

VERY IMPORTANT:

- Upload **e-copy** of your answers, on canvas.
- Time estimate: <u>**HEAVY: 15-30 hours**</u> (≈ 1 hour for Q1; 5-10 hours per programming question)

Reminders:

- *Plagiarism*: Homework is to be completed *individually*.
- *Typeset* your answers. Illegible handwriting may get zero points.
- *Late homeworks*: Follow the published policy

For your information:

- Explanations are *optional*, and will only be used to for partial credit, if the main answer is off.
- No need to provide your code.
- Graded out of **100** points; **4** questions total

Revision : 2024/11/22 22:56

Question	Points	Score
Density paradox	10	
Mean-median paradox	30	
Similarities and SVD	30	
Fourier	30	
Total:	100	

Question 1: Density paradox .... [10 points]

See Figure 1: Consider a cloud of N=1,000 2-d points uniformly distributed across the diagonal of the unit square. Also consider the point $\mathcal{P}$= (0.5, 0.5) in the middle. Compute the density of the cloud around $\mathcal{P}$.

Use squares of side s ($s = 1, 1/2, 1/4$), centered around $\mathcal{P}$. Define as:

- N_s: the estimated number of points inside the square
- $D(s) = N_s/s^2$: the density estimate

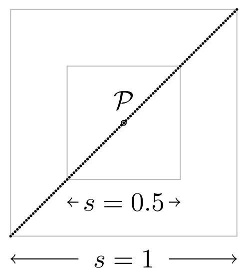


Figure 1: Compute the density around $\mathcal{P}$

.

(a) **[3 points]** Compute the density $D(1)$ for the square of side s=1 (centered at point $\mathcal{P}$)

(a) **N/1 = 1,000**

(b) **[3 points]** : Repeat for $D(0.5)$ (side $s = 0.5$)

(b) **2N = 2,000**

Solution: N/2 / s^2 = N/2 * 4 = 2N

(c) **[3 points]** : Repeat, for side $s = 1/4$

(c) **4N**

(d) **[0 points]** Why is the density fluctuating?

Solution: Because of the implicit assumption that the cloud of points is uniformly distributed (fractal dimension 2) in the 2-d space around $\mathcal{P}$. For points along a line (or along any other fractal) the density will grow, as the side of the square shrinks.

(e) **[1 point]** What should you do if this happens in a real dataset you are studying? (Mark all the answers that you agree with)

A. Compute the (local) fractal dimension instead

B. Fix the side s to a reasonable value.

C. Give the limit of the density $D(s)$ as $s \to 0$

D. none of the above

Grading info: *Any of A,B: full point*

Grading info: *full point if "A,B,C" although 'C' is wrong.*

Grading info: *'C' is wrong because the density tends to infinity with shrinking s*

Question 2: Mean-median paradox .... [30 points]

(*This is from a real story:*) Suppose you are interning at a financial institution, and your mentor would like you to study the statistics of the amounts of their financial transactions, and specifically the average.

The file `amounts1M.csv.gz` in the folder `http://www.cs.cmu.edu/~christos/courses/826.F24/HOMEWORKS/hw4-data/` mimics this scenario: It has $N = 2^{20}$ lines (plus a header line), each data line corresponding to a transaction, and each with a single number that stands for the amount of the transaction.

Write code to estimate the mean $m(n)$ for the first n transactions (n=2^{10}, 2^{11}, ..., 2^{20}=N); and also to estimate the median $\mu(n)$ for the same settings.

(a) **[10 points]** Give the plots for $m(n)$ and $\mu(n)$

Solution: The median should be stable, the mean should be wildly oscillating and growing. See Figure 2.

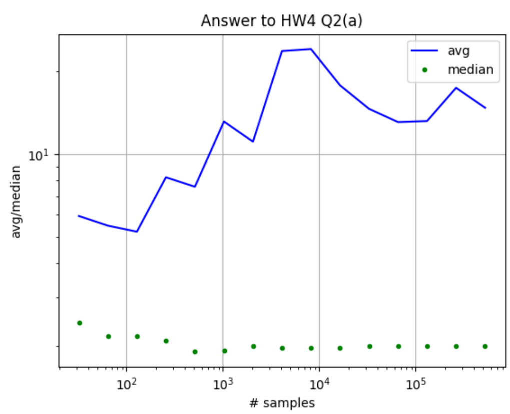


Figure 2: Mean and median, vs sample size.

(b) **[10 points]** You suspect that the data may follow a Pareto distribution. Plot the CCDF (= Prob($X \geq x$)) of the full dataset, in log-log scales.

Solution: See Figure 3

(c) **[5 points]** The CCDF plot seems to confirm that a Pareto distribution fits well - you conjecture that the formula is:

$$CCDF(x) \equiv \text{Prob}(X \geq x) = 1/x \quad (x \geq 1) \tag{1}$$

Compute the theoretical median μ, from Eq. 1.

(c) μ**=2**

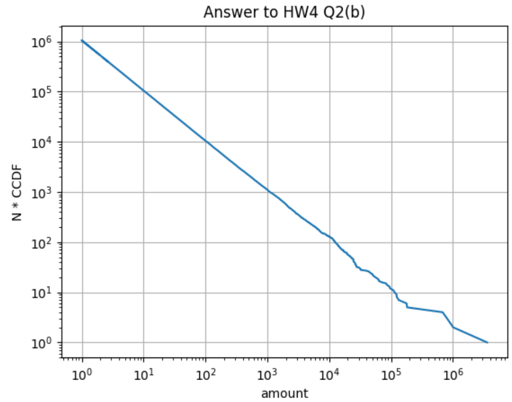


Figure 3: Plot for Q2(b): CCDF of given data (log-log scales).

Solution: $\text{CCDF}(\mu) = 1/2 = 1/\mu$.

(d) [**5 points**] Compute the theoretical average m, as the sample size $n \to \infty$

(d) $m = \infty$

Solution: $m = \int_1^\infty x \ \text{pdf}(x)dx$ and $\text{pdf}(x) = -\frac{\partial \text{CCDF}(x)}{\partial x}$

(e) [**0 points**] What would you report to your mentor?

A. Just report the average for the N=2^{20} given amounts

B. Suggest that you also compute the variance

C. Suggest a deep-dive, to make sure that the amounts follow a Pareto distribution

D. Suggest that you use the median instead

Grading info: 'A' is misleading, letting the mentor assume that the distribution is Gaussian

Grading info: 'B' would give very high variance - but it would still be misleading, silently implying a Gaussian distribution.

Question 3: Similarities and SVD ....[30 points]

(*Also based on (multiple) true stories:*) Suppose you have N=25 objects (eg., documents) and you are given all the cosine similarities in an $N \times N$ similarity matrix. Such a matrix is available as `similarities25.csv` at `http://www.cs.cmu.edu/~christos/courses/826.F24/HOMEWORKS/hw4-data/`

Clearly, the diagonal is all '1'. You want to visualize the dataset, and specifically, you want to find and scatter-plot the best 2-dimensional points on the unit circle, that will produce close similarity scores.

(a) **[10 points]** Do the SVD to the similarity matrix. What is its effective rank r? ('effective' means that you can ignore the small singular values, say ≤ 0.001)

(a) **2**

Solution: The two main singular values are: 4.73273065 1.61284226 - the rest are very small.

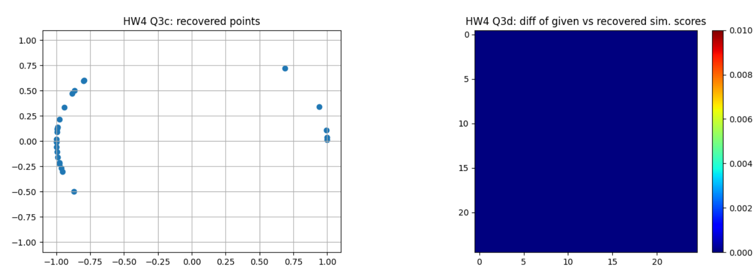
(b) **[5 points]** How many dimensions d would you need to find N=25 points that produce the same similarity matrix?

(b) **2**

<u>*Grading info:*</u> *-1pt, if '25' - that would be correct in the worst-case scenario, not in our case.*

(c) **[10 points]** Find N=25 points in 2-d, whose cosine similarities are as close to the original similarity matrix as possible; plot them as a scatterplot, like the one of Figure 5(a). Notice that there are multiple correct solutions (since rotation and mirroring preserve angles).

Solution: See Figure 4(c) below.

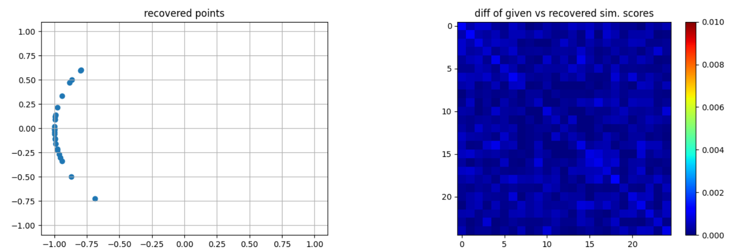


(c) recovered points (d) diff of similarities

Figure 4: Plot for Q3(c)-(d): Recovered points (left); difference of similarity matrices (right).

(d) [**5 points**] Plot (eg, `imshow()` of python) the $N \times N$ matrix of the differences between the original similarities, and the recovered ones - something like Figure 5(b)

Solution: See Figure 4(d)



(a) example of scatterplot (b) example of diff matrix

Figure 5: Visualizations: (a) of the recovered 2-d points; and (b) of the differences between original and recovered similarities

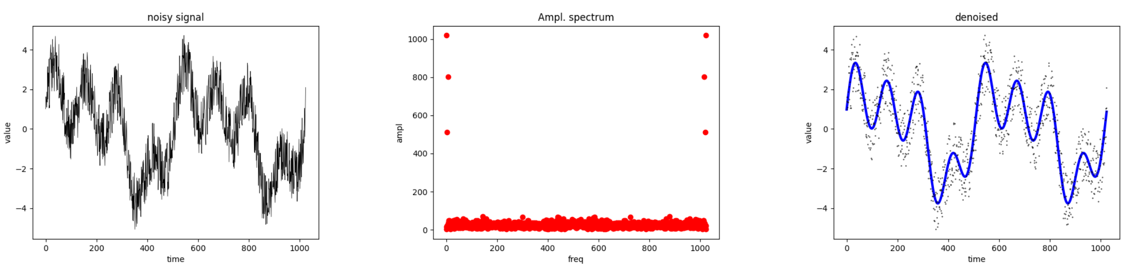
Question 4: Fourier ....[30 points]

(*Denoising is a popular task in signal processing*). Suppose you have a noisy time sequence $s(t)$ of duration N, like in Figure 6(a). It could be an intercepted intelligence signal; or the sales of a product over time; or the sound of a honeybee (main pollinators for our agriculture), or light from a star in astrophysics.

You suspect there are periodicities, of the form:

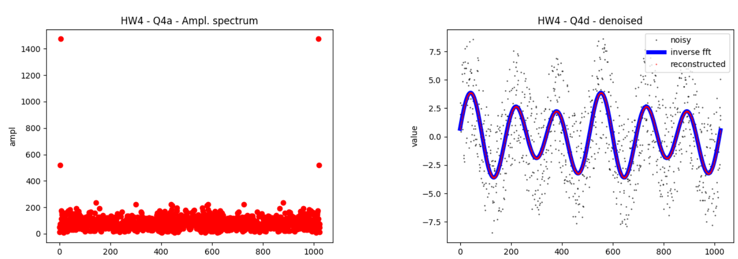
$$\begin{aligned} s(t) \;&=\; A_1 \sin(2\pi f_1 t/N + \phi_1) + \\ &\quad \ldots \\ &\quad A_k \sin(2\pi f_k t/N + \phi_k) + \\ &\quad noise(t) \end{aligned} \tag{2}$$

You want to find how many (k) such frequencies are there, as well as their parameters: Amplitudes A_i, frequenceies f_i, and phases ϕ_i ($i = 1, \ldots, k$)



(a) noisy signal (b) its Fourier spectrum (c) denoised signal

Figure 6: Example of a signal and its plots.



(a) Fourier spectrum (d) denoised signal

Figure 7: Answers: Q4a: Spectrum (left); Q4d: denoised signal (right).

The signal you are to work on is at `sounds1024.csv` in the folder `http://www.cs.cmu.edu/~christos/courses/826.F24/HOMEWORKS/hw4-data/`. It consists of 1024 lines, each having a value. There is no header.

(a) **[10 points]** Compute and plot the amplitude spectrum, as in Figure 6(b).

Solution: See Figure 7(a)

(b) **[5 points]** Spot the highest amplitude(s); how many are high, in your plot? (in the sample plot of Figure 6(b), one would argue that there are k=3 dominant frequencies).

(b) **2**

(c) **[10 points]** Give the amplitude, frequency, and phase, for each of the dominant components.

Solution: k= 2 dominant frequencies. The dominant FFT coefficients are The corresponding sinusoids of Eq 2 are:

- freq=4, Xreal=351.2042, Ximag=-380.1981
- freq=6, Xreal=1.7427, Ximag=-1474.1422

Translated to the parameters of the sinusoids, we have

- freq=4, Amplitude=1.01, $\phi = 42.72$ degrees
- freq=6, Amplitude=2.87, ϕ=0.06 degrees

FYI, the actual parameters were:

- $A_1 = 4$; f_1= 1, $\phi_1 = \pi/4 = 45°$
- $A_2 = 6$; f_1= 3, $\phi_2 = 0$

with noise amplitude: 10.

Grading info: 8/10pts if they only give the fft coefficients

Grading info: -1pt if they give amplitude+phase of Fourier transform, but do not give the A_i, ϕ_i of Eq.(2)

(d) **[5 points]** Drop all the other frequencies, and plot your de-noised signal (blue curve) along with the dots of the original signal as in Figure 6(c).

Solution: See Figure 7(d)

Hint: Use an existing library, like scipy fft. Any other package (matlab, octave, R), is fine.