Carnegie Mellon University
Department of Computer Science
15-826 Multimedia Databases and Data Mining
C. Faloutsos, Fall 2025

Homework 4 - Solutions

Due: **pdf, on canvas, at 2:00pm, on 11/21/2025**

VERY IMPORTANT:

- Upload **e-copy** of your answers, on canvas.
- Time estimate: **HEAVY: 10-20 hours** (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:

- No need to provide your code.
- Graded out of **100** points; **2** questions total

Revision : 2025/11/23 14:53

Question	Points	Score
Mystery dataset - fractals and SVD	60	
Discrete Fourier Transform (DFT)	40	
Total:	100	

Question 1: Mystery dataset - fractals and SVD .... [60 points]

Motivation: In multiple real-life settings, we are given a cloud of n points in d dimensions, and we want to get an idea of how the dataset looks like - are there redundancies? clusters? outliers? Do the points lie on a lower-dimensionality manifold (like, e.g, on the periphery of a circle)?

Thus, consider the 4-d 'mystery' dataset at `https://www.cs.cmu.edu/~christos/courses/826.F25/HOMEWORKS/hw4_data/mystery.bsv` The goal is to get an intuition about this dataset, using the tools we have learned: Fractals, SVD, and visualization.

(a) **[15 points]** The dataset may be self-similar. Thus, compute the fractal dimension of the dataset. We recommend that you use the code at `https://www.cs.cmu.edu/%7Echristos/SRC/fdnq_h_x11.tar` Check the README.fdnq file; it needs perl (v5) and x11. It was tested on the `unix.andrew.cmu.edu` machines.

The script gives the slope, y-intercept and correlation coefficient; give these numbers, up to 3 decimal digits.

Solution: with fdnq.pl -r0.1 -R10000:
slope= 1.472 y-intcpt= 7.212 corr= 0.998

(b) **[5 points]** We want to visually verify that the dataset is self-similar. Thus, please give the plot of the correlation integral (ie., use the '-v' verbose flag),

Solution: See Figure 1(a)

(c) **[15 points]** Find how many (if any) of the dimensions are redundant. That is, do (centered) SVD to the $n \times d$ data matrix, and give its singular values (up to 2 decimal digits)

Solution: 46450.63, 27200.33, 10.19, 9.82

(d) **[5 points]** How many singular values would you keep, to reconstruct the matrix?

(d) **2**

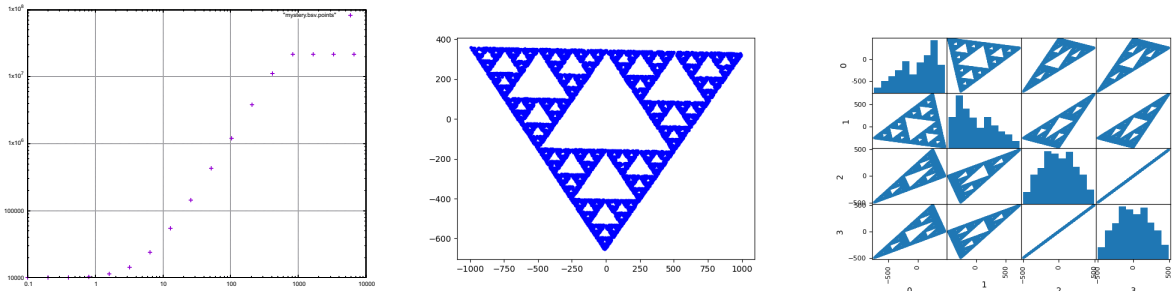
Solution: The two main singular values - the rest are very small.

(e) **[5 points]** Visualization: project the dataset on the two best axis (according to SVD), and provide the resulting 2-d scatterplot.

Solution: See Figure 1(b).

(f) **[5 points]** Give the (4-d) vectors of the above two best axis that you are projecting.

Solution: $v_1 = [0.4895, 0.5104, 0.4999, 0.4999]$
$v_2 = [-0.7144, 0.6996, -0.0074, -0.0074]$



(a) corr. int. (b) 2d projection (c) pair-plot

Figure 1: Plots from the mystery dataset.

The original ones were $v_1 = [0.5, 0.5, 0.5, 0.5]$ and $v_2 = [a, -a, 0, 0]$ where $a = \sqrt{(2)}/2 = 0.7071$

(g) **[5 points]** Which of the following (if any) matches your understanding of the 'mystery' dataset. That is, what can you say about the manifold that the points belong to:

A. A Sierpinski triangle

B. A Koch snowflake

C. A Cantor dust

D. A line or smooth curve

E. A 2-d plane (or smooth surface)

F. A 3-d hyper-plane

G. None of the above (elaborate, briefly)

(h) **[5 points]** Provide the pair-plot (= scatter-matrix) of the original 4-d 'mystery' dataset. *Hint:* Try `python pandas`: `scatter_matrix()`

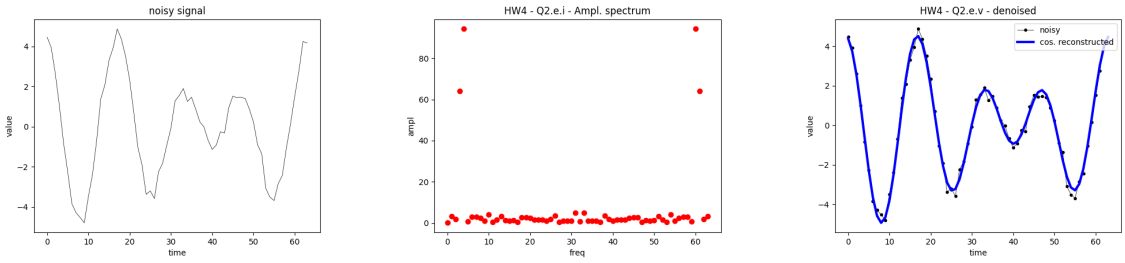
Solution: See Figure 1(c)

(i) **[0 points]** Does the pair-plot agree with your expectations?

Solution: Yes - it seems to be a Sierpinski triangle, on a 2-d plane, that is somehow rotated in the 4-d space.

Question 2: Discrete Fourier Transform (DFT) . . . . . [40 points]

Motivation: The goal is to illustrate how DFT can help with recovering the equations of a noisy signal, and thus, with the denoising of it. Noisy signals appear in multiple settings, like, e.g., sales of a product over time; electrocardiogram signals; weather temperature signals, and more.



(a) signal (b) Ampl. spectrum (c) denoised

Figure 2: original, DFT and denoised 'mysteryWave64' signal.

Hint: Use an existing library, like scipy fft. See the tutorial at scipy 1-D DFT. We shall follow the tutorial, and use its definition for the N-point DFT ($j \triangleq \sqrt{-1}$):

$$X_f = \sum_{t=0}^{N-1} x_t * \exp(-2\pi j f t/N) \quad f = 0, \ldots, N-1 \tag{1}$$

and for the inverse:

$$x_t = 1/N \sum_{f=0}^{N-1} X_f * \exp(2\pi j f t/N) \quad t = 0, \ldots, N-1 \tag{2}$$

(Reminder: these definitions only differ by a $\sqrt{N}$ factor from the ones in the course slides.)

The over-arching assumption is that our N-point signal consists of k frequencies expressed as cosine waves, and some noise:

$$\begin{aligned} x_t \quad = \quad & A_1 \cos(2\pi f_1 t/N + \phi_1) + \\ & \ldots \\ & A_k \cos(2\pi f_k t/N + \phi_k) + \\ & noise(t) \end{aligned} \tag{3}$$

where A_m, f_m and ϕ_m are, respectively, the amplitude, frequency and phase of the m-th cosine wave. The goal of the exercise is to illustrate how the DFT coefficients can help us recover the parameters of the cosine waves.

Specifically, we will use the 'mysteryWave64.csv' file at `https://www.cs.cmu.edu/~christos/courses/826.F25/HOMEWORKS/hw4_data/mysteryWave64.csv` with N=64 lines, representing a signal s_t $t = 0, \ldots, N$. Figure 2 gives a plot of it.

We start with some easy cases, then derive the equations for amplitude and phase, and then proceed to denoise the "mysteryWave64"signal itself.

(a) (Warm-up - step1 - plain cosine) Generate a pure cosine of N=32 points, with frequency f=1, amplitude A=1 and phase ϕ=0, that is $x_t = 1 * cos(2\pi 1t/N + 0)$ $(t = 0, \ldots, N-1)$; report the non-zero coefficients of its DFT. Do count the complex conjugates, that is, if only $X_3 \neq 0$ and $X_{N-3} = X_3^* \neq 0$, then your answer would be '2'.

i. **[1 point]** How many are the non-zero coefficients?

i. **2**

ii. **[3 points]** Give their values

Solution: $X_1 = X_{N-1}^* = 16$

(b) (Warm-up - step 2 - shifted cosine) Repeat, when the above cosine wave is shifted, that is, $\phi = \pi/3 = 60°$ (and all else unchanged: f=1, A=1, N=32) - that is $x_t = 1 * cos(2\pi 1t/N + \pi/3)$ $(t = 0, \ldots, N-1)$; report the non-zero coefficients of the DFT:

i. **[1 point]** How many are the non-zero coefficients?

i. **2**

ii. **[3 points]** Give their values

Solution: $X_1 = X_{N-1}^* = 8 + 13.856j$

(c) (Warm-up - step 3 - longer cosine) Repeat, for a longer cosine wave ($N = 64$), with A=1, f=1, $\phi = \pi/3$. that is $x_t = 1 * cos(2\pi 1t/64 + \pi/3)$ $(t = 0, \ldots, 64-1)$; report the non-zero coefficients of the DFT:

i. **[1 point]** How many are the non-zero coefficients?

i. **2**

ii. **[3 points]** Give their values

Solution: $X_1 = X_{N-1}^* = 16 + 27.713j$

(d) Consider a pure cosine wave of N-points, $x_t = A*cos(2\pi ft/N+\phi)$ $(t = 0, \ldots, N-1)$; and let its two non-zero DFT coefficients be: $X_f = R+Ij$ and $X_{N-f} = X_f^* = R-Ij$. We want to express the amplitude A and phase ϕ as functions of the real part R and imaginary part I of the X_f coefficient.

i. **[4 points]** Pick, or write-in, the correct formula for the amplitude A

A. $A = \sqrt{R^2 + I^2}$

B. $A = 2 * \sqrt{R^2 + I^2}$

C. $A = N * \sqrt{R^2 + I^2}$
D. $A = 2 * N * \sqrt{R^2 + I^2}$
E. $A = 2/N * \sqrt{R^2 + I^2}$
F. $A = 2/\sqrt{N} * \sqrt{R^2 + I^2}$
G. $A = 1/\sqrt{N} * \sqrt{R^2 + I^2}$
H. None of the above (give correct formula)

ii. [**4 points**] Choose, or write-in, the correct formula for the phase ϕ:

A. $\phi = \tan(I/R)$
B. $\phi = \tan(R/I)$
C. $\phi = \arctan(I/R)$
D. $\phi = \arctan(R/I)$
E. $\phi = 2 * I/R$
F. $\phi = 2 * R/I$
G. $\phi = 2 * \tan(R/I)$
H. $\phi = 2 * \tan(I/R)$
I. None of the above (give correct formula)

(e) As mentioned earlier, we want to denoise the 'mysteryWave64.csv' file at `https://www.cs.cmu.edu/~christos/courses/826.F25/HOMEWORKS/hw4_data/mysteryWave64.csv` with N=64 lines, representing a signal s_t $t = 0, \ldots, N$. The signal is a mixture of some cosine waves, plus noise (see Figure 2)

i. [**4 points**] Plot the amplitude spectrum of the DFT

Solution: See Figure 2(b)

ii. [**4 points**] Ignore small values - how many are the main coefficients?

ii. **4**

iii. [**4 points**] Give all those coefficients (eg., $X_3 = 32 + 5j$, etc)

Solution: $X_3 = X^*_{61} = 46.025 + 44.437j$
$X_4 = X^*_{60} = 94.232 - 3.097j$

iv. [**4 points**] Compute and report the amplitude A_f and phase ϕ_f of those main frequencies (eg., $A_3 = 12.4$, $\phi_3 = \pi/2 = 90°$) Report the phases in degrees (`math.degrees(math.pi)` = 180°).

Solution: freq=3, Amplitude=1.999, math.degrees(phase)=43.995
freq=4, Amplitude=2.946, math.degrees(phase)=-1.882
Grading info: no penalty, if A3 = 63.97, A4=94.28

v. [**4 points**] For visual verification, plot (a) the 'mysteryWave64.csv' signal, superimposed on (b) your reconstruction of it (using Eq 3 with the appropriate cosine waves, and zero noise)

Solution: See Figure 2(c)
Grading info: [-1] if the reconstruction is 2x than the correct one.