EAS 502, Fall 2017, Midterm Exam, Part II
Take Home, Due 12 PM EST Nov. 13, 2017, submitted to UBLearns
Total Possible Points: 75

Work **all** problems. Please read all directions carefully. You may use course notes, books, internet search, and tools such as Matlab or Python. If you use external references such as journal papers, be sure to cite them accordingly. **Consultation, either in-person or virtually, with anyone other than Professor Bauman or the TAs is STRICTLY FORBIDDEN.** All work submitted must be your own! Submit all code written and one PDF document with your answers to each problem. Your PDF document should be well organized and your code well commented. Any copying of work or consultation with others will result in a grade of **ZERO** for the **ENTIRE** midterm exam.

1. (Linear Regression, 15 pts.) A planet follows an elliptical orbit, which can be represented in a Cartesian (x, y) coordinate system:

$$ay^2 + bxy + cx + dy + e = x^2$$

You are given the following observations of the planet's position:

x	1.02	0.95	0.87	0.77	0.67	0.56	0.44	0.3	0.16	0.01
y	0.39	0.32	0.27	0.22	0.18	0.15	0.13	0.12	0.13	0.15

   (a) (7 pts.) Formulate a linear regression problem to determine the coefficients a, b, c, d, e from the observations.

   (b) (5 pts.) Develop a program (in Matlab or Python, your choice) to compute the coefficients from the data.

   (c) (3 pts.) Plot the data points and the fit curve on the same graph.

2. (Numerical Rank, 30 pts.) It turns out that in the presence of floating point error, the rank of a matrix is somewhat meaningless. Consider the following theorem: Let $A \in \mathbb{R}^{m\times n}$, with $\text{rank}(A) = r < \min(m, n)$. Then, for every $\varepsilon > 0$, there exists a full rank matrix $A_\varepsilon \in \mathbb{R}^{m\times n}$ such that $\|A - A_\varepsilon\| < \varepsilon$. This means for any rank-deficient matrix, there is a full rank matrix arbitrarily "close" to it.

   (a) Consider

$$A = \begin{bmatrix} 1/3 & 1/2 & 2/3 \\ 2/3 & 2/3 & 4/3 \\ 1/3 & 2/3 & 3/3 \\ 2/5 & 2/5 & 4/5 \\ 3/5 & 1/5 & 4/5 \end{bmatrix}$$

      i. (2 pts.) What is the exact rank of this matrix? Justify. (Hint: Look at the columns)

      ii. (5 pts.) Compute the SVD and discuss how you would assign the rank, taking into account double precision floating point error.

      iii. (3 pts.) Consider a 2000×2000 matrix with singular values $\sigma_n = (0.9)^n$. How would you assign a rank to this matrix?

   (b) It turns out that Problem 1 is close to rank deficient. We will repeat it now, but study the effect of noise.

      i. (10 pts.) Repeat Problem 1, but now adding noise to each coordinate. Use uniformly random noise distributed on the interval $[-0.005, 0.005]$. Solve the regression problem again with the noisy data to obtain new coefficients. Compare the new values and the previous values. What effect do you see on the plot of the orbit? Explain.

ii. (10 pts.) Now repeat solving the problem, for both the original and the noisy data, using a low rank approximation based on the SVD. Experiment with cutoff tolerances of say $10^{-k}, k = 1, \ldots, 5$. Compare the solutions for each. How well do the resulting orbits fit the data points as the tolerance and rank vary? Which solution would you regard as better: one that fits the data more closely, or one that is less sensitive to small perturbations in the data? Why?

3. (SVD, 30 pts.)

In this problem, we study the voting patterns of the United States congress using the SVD. Data for the 2012 congressional voting record, both the House and the Senate, is included with this assignment (and it was pulled from the web). For each there is a `names.txt` file, a `parties.txt` file, and a `votes.txt` file containing the names, the political party, and the votes for all of 2012 for each congress member. The voting record data is essentially in a matrix structure, with each row corresponding to a particular congress member and each column as a different vote by that congress member. The political party code is 100 for Democrats, 200 for Republicans, and then any other code we will label as Independent. In the voting data, a value of 1 corresponds to "Yea", a value of -1 corresponds to "Nay", and 0 is abstention (via absence or otherwise).

(a) (4 pts.) Write a Matlab script that parses the Senate data using the `textscan` command (for the `names.txt` file), `load` command (for the `votes.txt` and `parties.txt` files), and `find` (for extracting party affiliation). You can alternatively use Python if you wish, but this should only be about 8 lines of Matlab.

(b) (3 pts.) Compute the SVD of the voting record and plot the singular values. What do you observe?

(c) (5 pts.) Create a scatter plot of the first and second columns of V. That is, each coordinate in the scatter plot will be $(v_1(j), v_2(j))$. Color each coordinate by party affiliation (Democrat, Republican, or Independent).

(d) (5 pts.) Examining this plot, what do you think each coordinate represents? That is, what do you think v_1 and v_2 are capturing? Consider generating other plots involving v_1 and/or v_2 to aid in understanding. (Hint: Think about the transformation of a unit circle by the SVD.)

(e) (5 pts.) Use a low rank approximation of the voting record using the first two dominant modes of the SVD. Based on the sign of each value in this approximation, assign a "Yea" or "Nay" vote and compare with the actual voting record. Count the total number of matches and compute the fraction of correct voting predictions based on the low rank approximation. Plot this number for each representative versus the v_1 vector. What do you observe? It can be instructive to plot the name of each representative on their respective point.

(f) (3 pts.) Repeat this exercise for the House data.

(g) (5 pts.) Based on this analysis, what can we conclude about the voting patterns in the United States Congress in the year 2012?