Midterm Exam, Part II : EAS 596

Shubham Sharma, Person No.: 50290293

November 5, 2018

Question 1(a):

The given equation, $ay^2 + bxy + cx + dy + e = x^2$ will be satisfied by the given data points (x_i, y_i) $\forall i \in [1, 10]$. This gives us a system of 10 equations which can be solved for a, b, c, d and e by writing this system of equations in the form of Ax = b .

Thus, the linear regression problem is formulated as follows:

$$\begin{bmatrix} y_1^2 & x_1y_1 & x_1 & y_1 & 1 \\ y_2^2 & x_2y_1 & x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{10}^2 & x_{10}y_{10} & x_{10} & y_{10} & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix} = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_{10}^2 \end{bmatrix}$$

Question 1(b):

The coefficients obtained from the data are as follows: a = -2.6356, b = 0.1436, c = 0.5514, d = 3.2229 and e = -0.4329

Question 1(c): The data points and the fit curve are plotted in Figure 1.

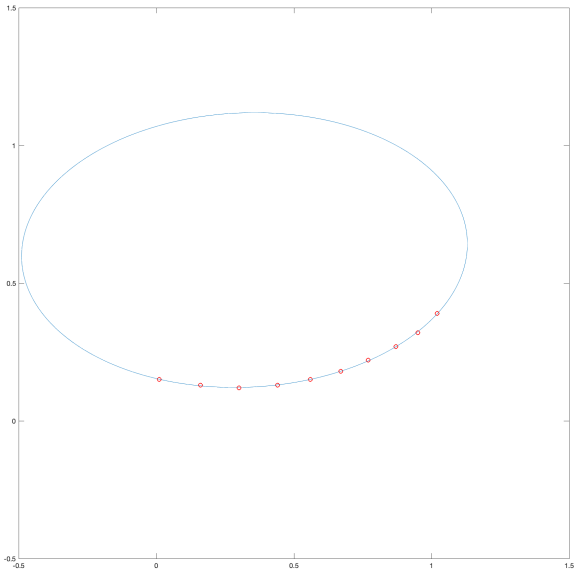

Figure 1: Question 1(c)

Question 2(a):

i. The exact rank of the given matrix A is 2 as computed using the *rank* command. This is because there are two independent columns in the matrix and the third column is the sum of the first two columns. In MATLAB, the *rank* command is used to calculate the rank by computing the number of singular values that are larger than a tolerance which is $max(size(A)) * eps(norm(A))$ by default.

ii. All floating point numbers cannot be represented exactly in binary form. Due to this, each arithmetic operation is generally affected by round off error. Basically, we can store only a subset of real numbers which are called floating point numbers since digital computers use a finite number of bits. Thus, on computing SVD we get three non-zero singular values where the third singular values is 1.926124×10^{-16}. The value of the default tolerance for the *rank* command comes out to be 2.2204×10^{-15} for the given matrix A. Thus, it ignores the last singular value which is less than the tolerance to compute the rank.

If we take the values of the matrix upto four decimal points, we observe the value of the rank to be 3 using the *rank* command. This is because the double precision floating point error is more here as compared to the previous case to represent the numbers. Because of this the third singular value comes out to be 7.960881×10^{-5} which is more than the tolerance level. Although, the rank should be 2 as we have seen in the i part and thus we need to increase the tolerance level to get the correct rank using the *rank* command.

iii. In double precision, the precision as determined by the mantissa is 2^{-52} which is roughly 2.22×10^{-16} which means the gaps between adjacent numbers are never larger than 2^{-52} in a relative sense. Thus, using the *rank* command, the rank comes out to be 276 for the matrix S. If we increase the number of significant figures to represent the values, we observe that the rank returned is 700 and also the condition number increases. Thus, we see as we increase the number of significant figures the round off error decreases and we get a more accurate rank. As already discussed, it ignores the values which are less than the tolerance.

iv. A $\in R^{m \times n}$

Rank of A is r which is less than min(m,n), thus for every $\epsilon > 0, there exists a full rank matrix A_\epsilon$

which is in $R^{m \times n} such that$

norm of (A-A_ϵ)$is less than \epsilon$.

Thus, we can say that in a matrix of m x n, there will be an upper bound on the singular values which in turn will depend on finite precision. Thus we get norm of (A-A_ϵ)$is less than \epsilon$.

Question 2(b):

i. The coefficients obtained from the data are as follows after setting a seed for the uniform random noise generated: a = -3.9069, b = 0.9126, c = 0.4486, d = 3.0819 and e = -0.3897.

The data points and the fit curve are plotted for the original as well as the noisy data.

We observe that the ellipse with noise vary a lot ranging from smaller than the original ellipse to very large as compared to the original ellipse. This is because all the points being considered lie on one side of the ellipse and thus very small changes in the data points lead to large changes in the curvature of the ellipse. The value of coefficients vary a lot as different random noise is generated and there is no particular trend seen.

ii. On observing the plots we see that in both cases, as the rank increases(the tolerance decreases), the curve becomes better in terms of its elliptical shape. When the tolerance is 10^{-1} we get rank: 3 and the shape of the curve is a hyperbola, for tolerance = 10^{-2} the rank obtained is 4 and the curve is not close to ellipse. As tolerance is decreased further, all singular values are included and hence we achieve full rank of 5: the shape of the curve is an ellipse. The same behaviour is observed for noisy as well as original data.

The solution which fits the data more closely tends to overfit and will not do a good job to make predictions on new data. The solution that is less sensitive to small perturbations in the data(noise)

can be too rigid sometimes and may have a high bias towards the training data i.e. the observations available to us. Thus, a balance is needed between the two to obtain the optimal solution.

Question 3:

b: The first two singular values for the matrix of senate votes comes out to be 144.1323 and 111.1622 which are the dominant ones over the other. This is also evident in the plot.

d: u_1 and u_2 depict the direction of the votes of both the parties. In other words, these vectors correspond to the principal components in the singular value decomposition. The vectors u_1 and u_2 are the directions of the first two principle semi-axes of an ellipse. They tell us the directions of the first two dominant singular value which is also evident in the divide which is shown in the plot between the votes of Republic and Democrats.

Although, there is a clear divide between the Replicans and the Democrats people like OBAMA, ENSIGN, SCHATZ, KIRK, HELLER etc. do not follow the partition exactly. SANDERS who is independent has a voting pattern similar to democrats. On trying with different plots, like u_1 with u_3, u_2 with u_4, u_2 with u_3 and u_6 with u_8 we find that the divide diminishes and the data becomes mixed. Since, the first two singular values are dominant, this behaviour is seen.

e: 41825 out of 48600 have been correctly predicted as per our computation which gives us an accuracy of 86% i.e. 86% of data is captured by rank 2 low matrix approximation. This is because the first 2 singular values are very dominant. From the plot of (U1 vs correct votes count) we can see the earlier partition by the parties. The people who were lying in between like OBAMA, ENSIGN, SCHATZ, KIRK, HELLER etc have very bad accuracy in the low rank approximation compared to others. Thus, it is difficult for the model to capture their data accurately in the low rank approximation.

f: Similar trend is seen in the house data as well.

g: It is clear that votes have been partitioned by parties. Democrats and Republicans follow their own patterns of voting and the only independent candidate: SANDERS followed the voting pattern of democrats. Very few people like OBAMA, ENSIGN, SCHATZ, KIRK, HELLER etc broke this partition and voted.