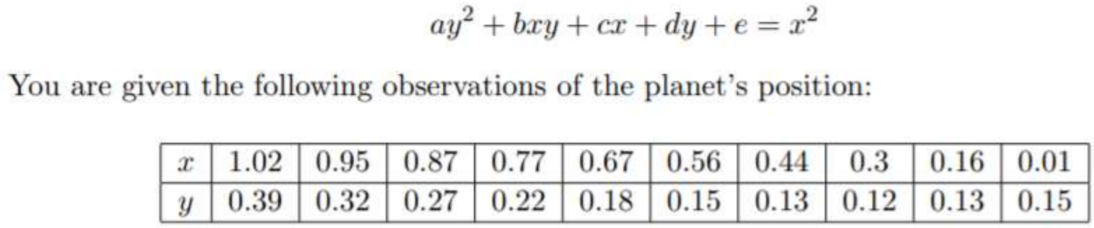
EAS596- Take Home Mid-Term

Paul M Girdler

Question 1

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

$$\boldsymbol{ay^2 + bxy + cx + dy + e = x^2}$$

$$\boldsymbol{M} = [y^2 + xy + cx + dy + 1]$$

$$\boldsymbol{Mp = b}$$

Where,

$$\boldsymbol{b} = x^2$$

$$\boldsymbol{p} = [a\ b\ c\ d\ e]$$

Now, we must solve for **p** using the pseudo inverse:

$$\boldsymbol{p = M^{+}b}$$

This will be the LS solution.

(b) Develop a program in Matlab to compute the coefficients from the data.

See Matlab script.

NOTE: I used the quadratic expression (15) from Wolfram Mathworld in the actual Matlab script.

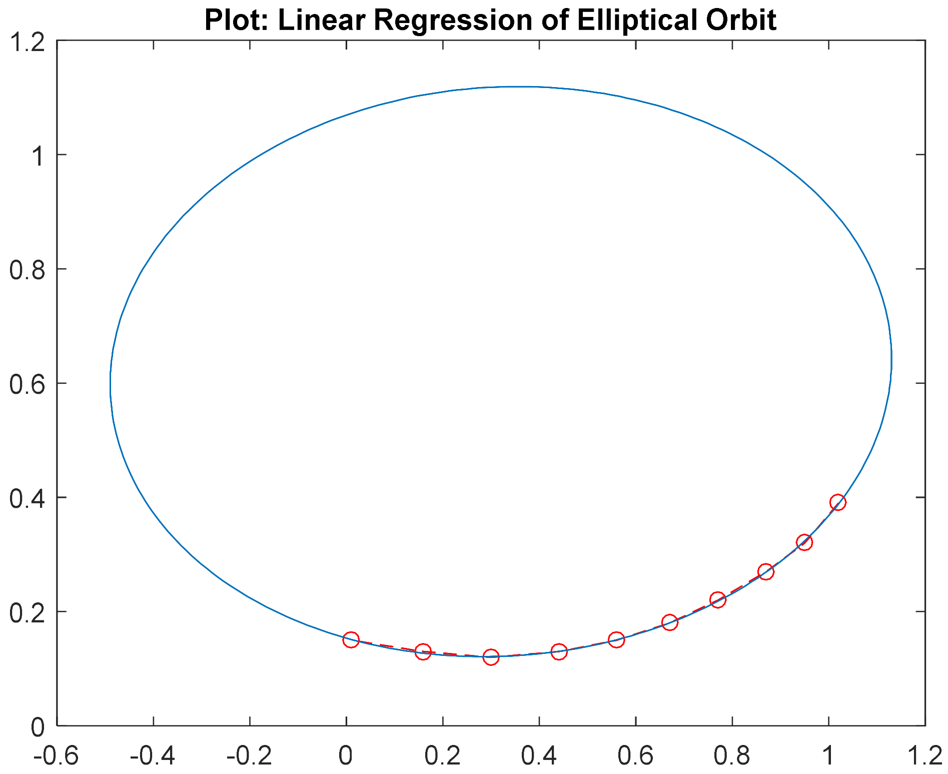
$$ax^2 + 2bxy + cy^2 + 2dx + 2fy + g = 0$$

This notation related better to other functions and literature related to elliptical functions.

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

TABLE – CALCULATED COEFFICIENTS

a	b	c	d	e
2.636	-0.072	-0.551	-3.223	0.433



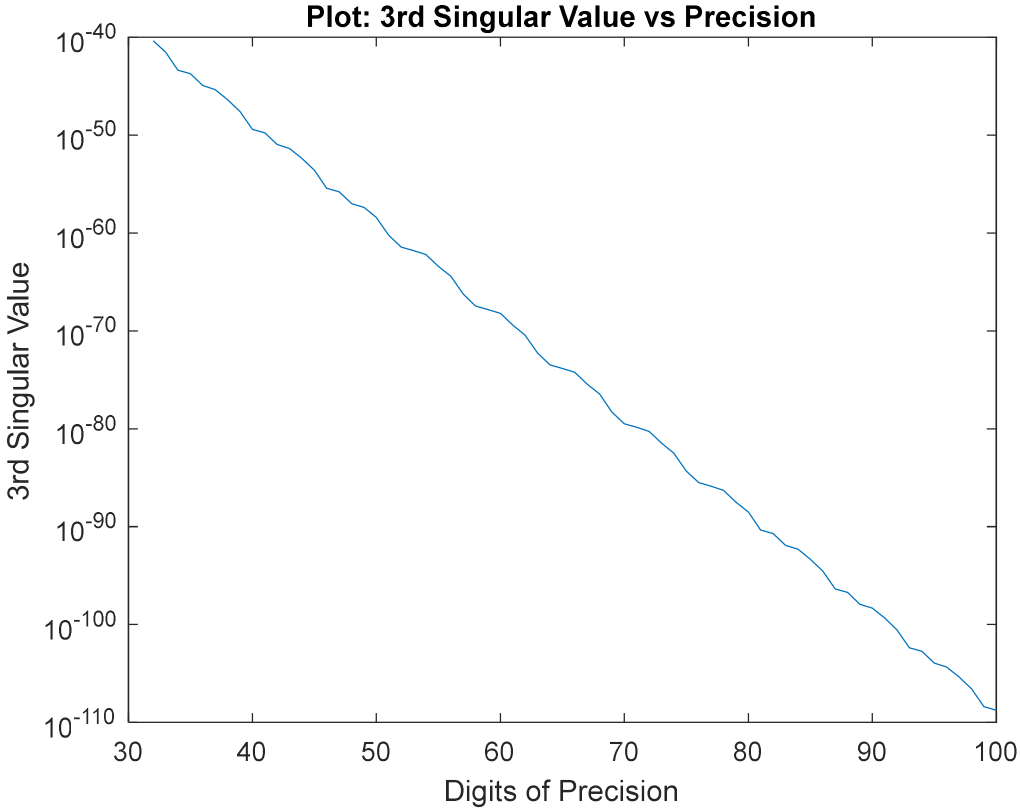
Question 1

(a) **It turns out that in the presence of floating point error, the rank of a matrix is somewhat ambiguous. Consider matrix A.**

(i) What is the exact rank of this matrix? Justify

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

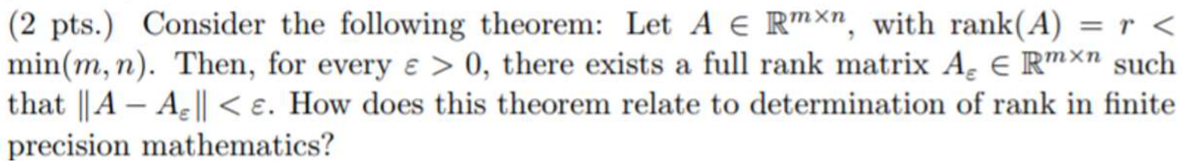
The exact rank of this matrix is 2. Using symbolic computation in Matlab gives the exact rank. If we decompose A into SVD form we see three singular values, implying rank = 3. However, as the number of digits of precision approaches ∞ we can see that σ_3 approaches 0. This means the rank will converge on 2, as σ_3 converges to 0.



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

This matrix is full rank, r = 2000. In both single and double precision $\sigma_{2000} \neq 0$, hence this is matrix is <u>still full rank</u>. However, if one was to approximate the rank for other purpose it would be best to weigh up the trade-off between the accuracy required for the problem against the computation expense.

(iv) Consider:



(2 pts.) Consider the following theorem: Let $A \in \mathbb{R}^{m\times n}$, with 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 $\|A - A_\varepsilon\| < \varepsilon$. How does this theorem relate to determination of rank in finite precision mathematics?

$$\left||\boldsymbol{A} - \boldsymbol{A}_{\boldsymbol{v}}|\right| < \boldsymbol{\varepsilon}$$

$$\left||\boldsymbol{A} - \boldsymbol{A}_{\boldsymbol{v}}|\right|^{2} < \boldsymbol{\varepsilon}^{2}$$

Now,

$$\left||\boldsymbol{A} - \boldsymbol{A}_{\boldsymbol{v}}|\right|^{2} = \sigma_{\boldsymbol{v}+1}^{2}$$

$$\sigma_{\boldsymbol{v}+1} < \boldsymbol{\varepsilon}$$

This means as the finite precision ε approaches 0, v will converge on r the rank of matrix A. This can be used to overcome the limitations of finite precision and calculate the exact rank of a 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) **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.**

TABLE – COMPARISON OF COEFFICIENTS

	a	b	c	d	e
(x,y)	2.636	-0.072	-0.551	-3.223	0.433
(x,y) + noise	2.801	-0.226	-0.501	-3.002	0.392
Relative error	-6.26%	-214.53%	9.06%	6.87%	9.33%

As evident in the table above introducing noise (perturbation) into the problem increases the error associated with computing the coefficients by regression. This increase the error associated with the LS prediction of the orbit as evident in the plot below.

(ii) **Now repeat solving the problem, for both the <u>original</u> and the <u>noisy data</u>, using a low rank approximation based on the SVD. Experiment with cutoff tolerances of 10^{-k}, k = 1 to 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?**

TABLE – COMPARISON OF COEFFICIENTS FOR ORIGINAL (NO NOISE)

tol	a	b	c	d	e
1.00E-01	-0.118	-0.497	-0.504	-0.805	0.167
1.00E-02	0.188	0.585	-0.752	-3.238	0.487
1.00E-03	2.636	-0.072	-0.551	-3.223	0.433
1.00E-04	2.636	-0.072	-0.551	-3.223	0.433
1.00E-05	2.636	-0.072	-0.551	-3.223	0.433

As we can see from the table above the solution is only stable for tolerances below 1E-3. This is also evident in the plot below, where the two LS solutions for tolerances equal to 1E-1, and 1E-2 appear as lines projecting into space.

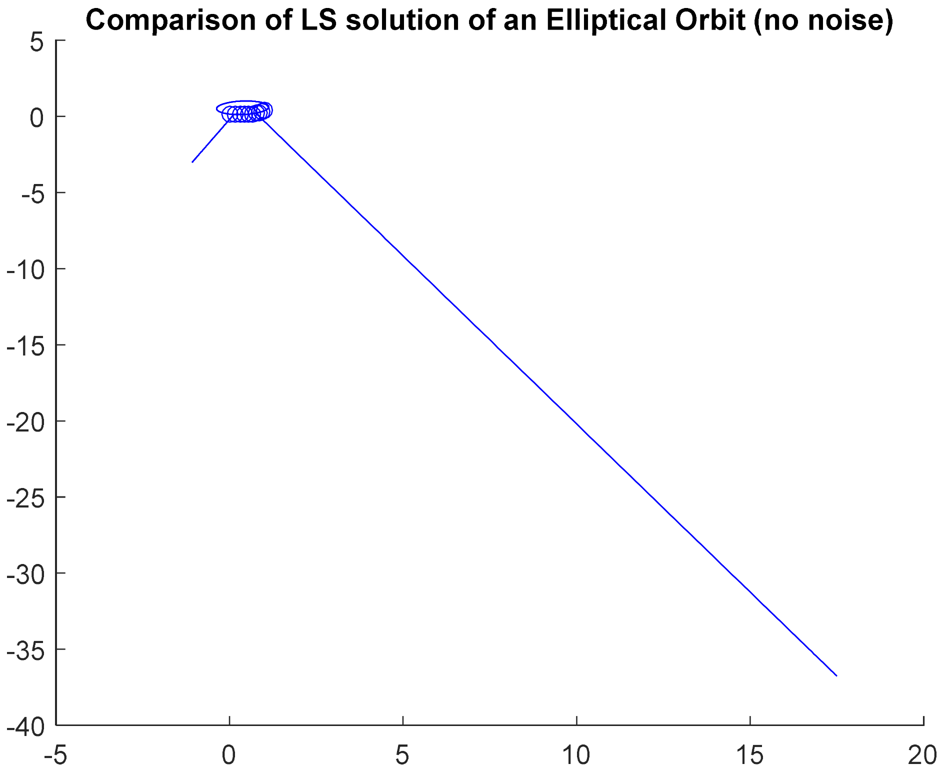
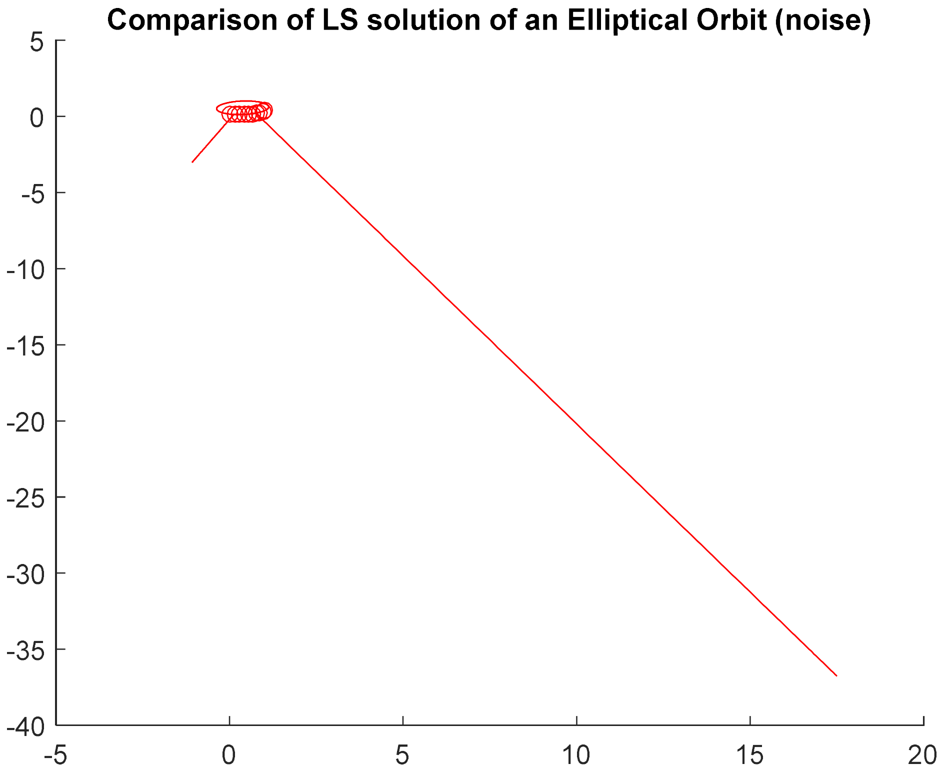


Table – Comparison of Coefficients (noise)

tol	a	b	c	d	e
1.00E-01	-0.119	-0.501	-0.495	-0.818	0.164
1.00E-02	0.157	0.480	-0.714	-3.019	0.449
1.00E-03	2.801	-0.226	-0.501	-3.002	0.392
1.00E-04	2.801	-0.226	-0.501	-3.002	0.392
1.00E-05	2.801	-0.226	-0.501	-3.002	0.392

Again we can see from the table above the solution is only stable for tolerances below 1E-3. This is also evident in the plot below, where the two LS solutions for tolerances equal to 1E-1, and 1E-2 appear as lines projecting into space.



This demonstrates how the conditioning of a matrix can make the LS solution sensitive to perturbations (noise).

Now, the generalized form of condition number is:

$$K(\boldsymbol{A}) = \left||\boldsymbol{A}|\right|\left||\boldsymbol{A}^{+}|\right|$$
$$K(\boldsymbol{A}) = \frac{\boldsymbol{\sigma_1}}{\boldsymbol{\sigma_n}}$$

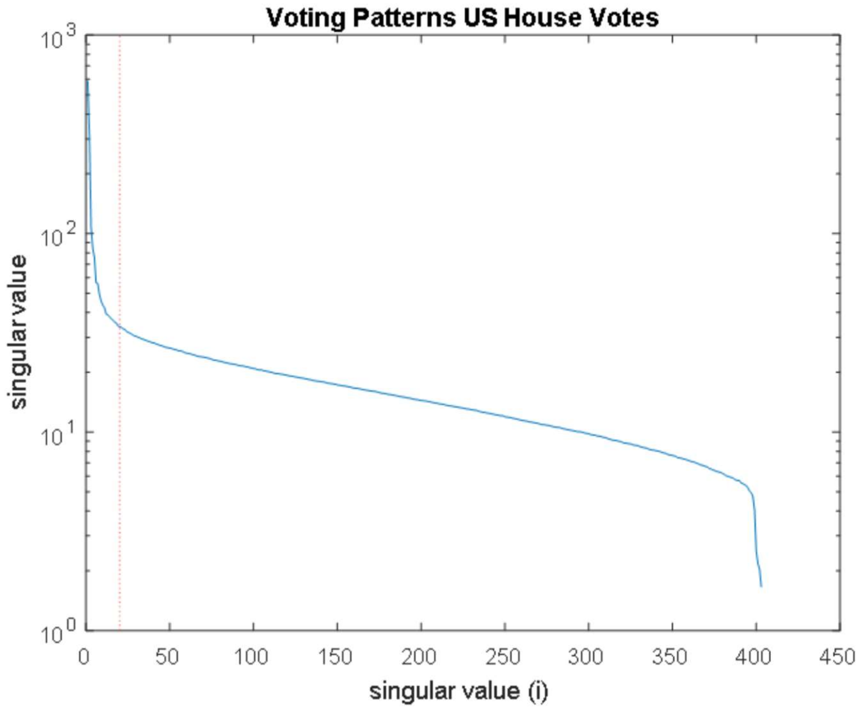
The tolerance is increased Matlab treats singular values of A that are smaller than the tolerance as zero. If $\boldsymbol{\sigma_n} < \frac{1}{\varepsilon}$ this is the same as having a large condition number. For this problem it appears that the threshold of stability appears to be at tolerances less than 1E-3.

Question 3

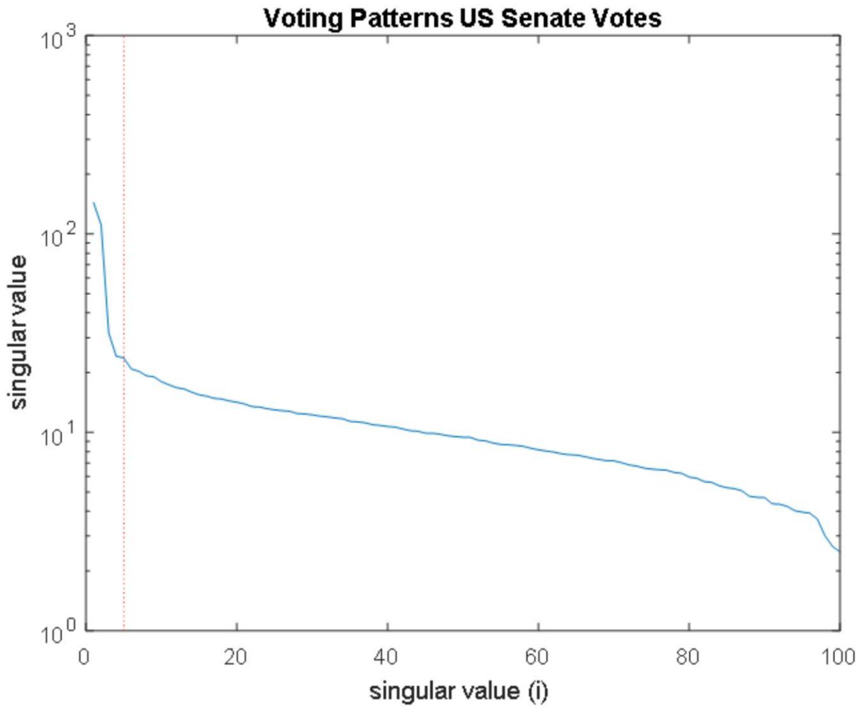
(a) Write a Matlab script that parses the Senate data. For example, you may use the textscan command (for the names.txt file), load command (for the votes.txt and parties.txt files), and find (for extracting party affiliation) in Matlab.

See Matlab code.

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

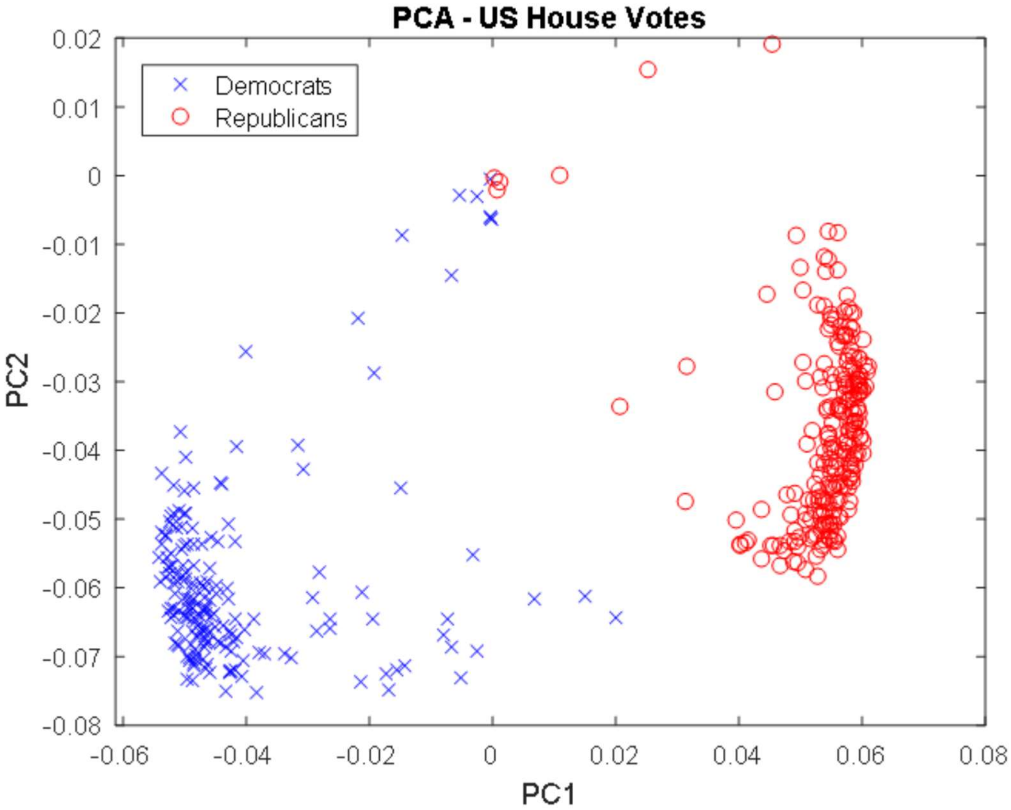


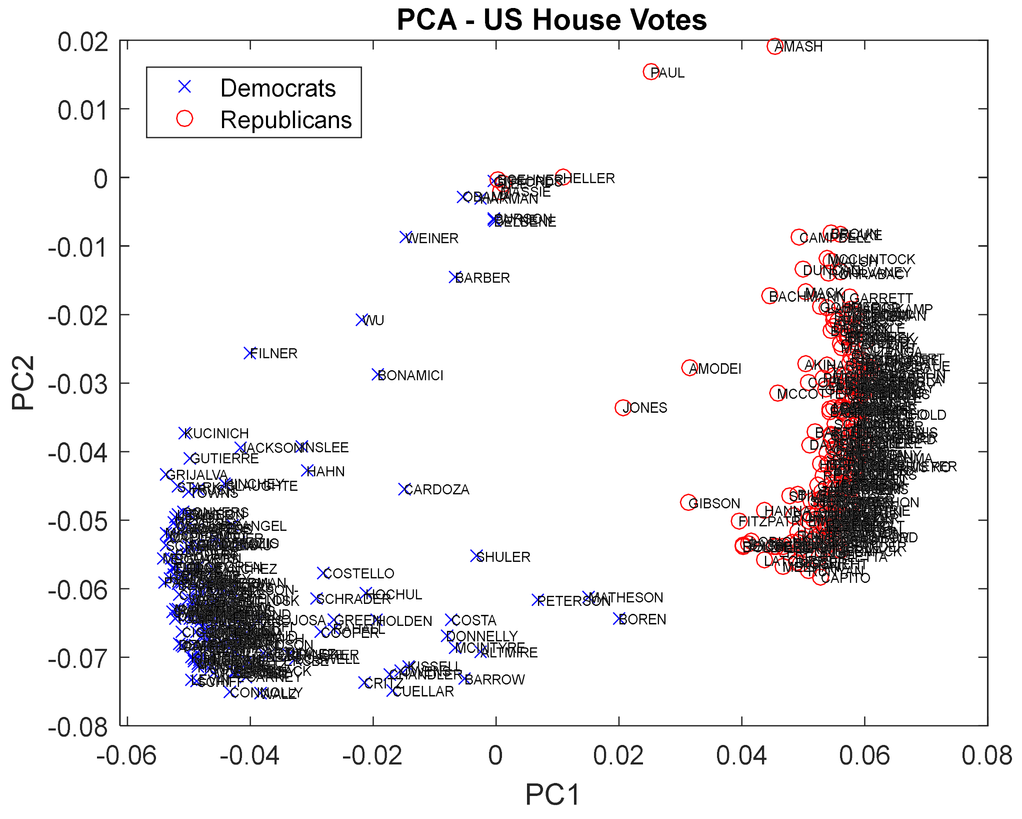
The SVD of the US House Voting Pattern was computed as required. The singular values were plotted. The plot above shows a number of dominant principle components between σ_1 and σ_{20}. The magnitude of the principle component decays sharply from σ_1 and σ_{20}. After σ_{20} the principle components decay at an exponentially constant rate. This trend continues until the last few principle components, where there is a sharp drop off.

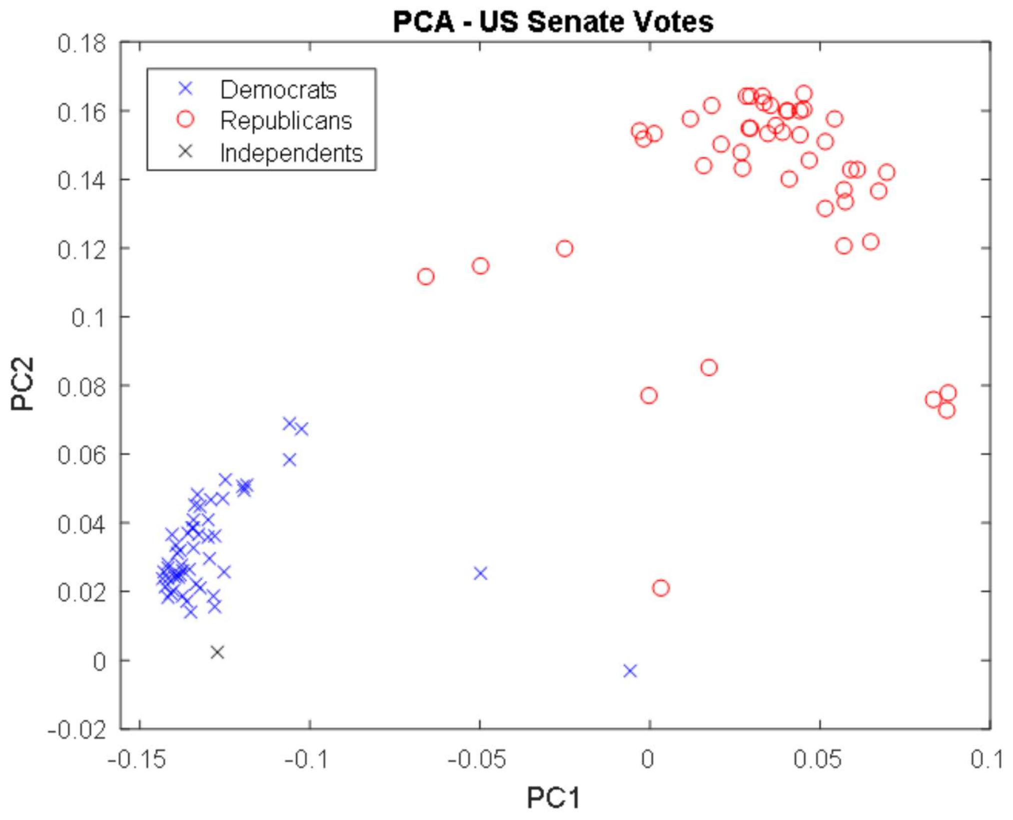


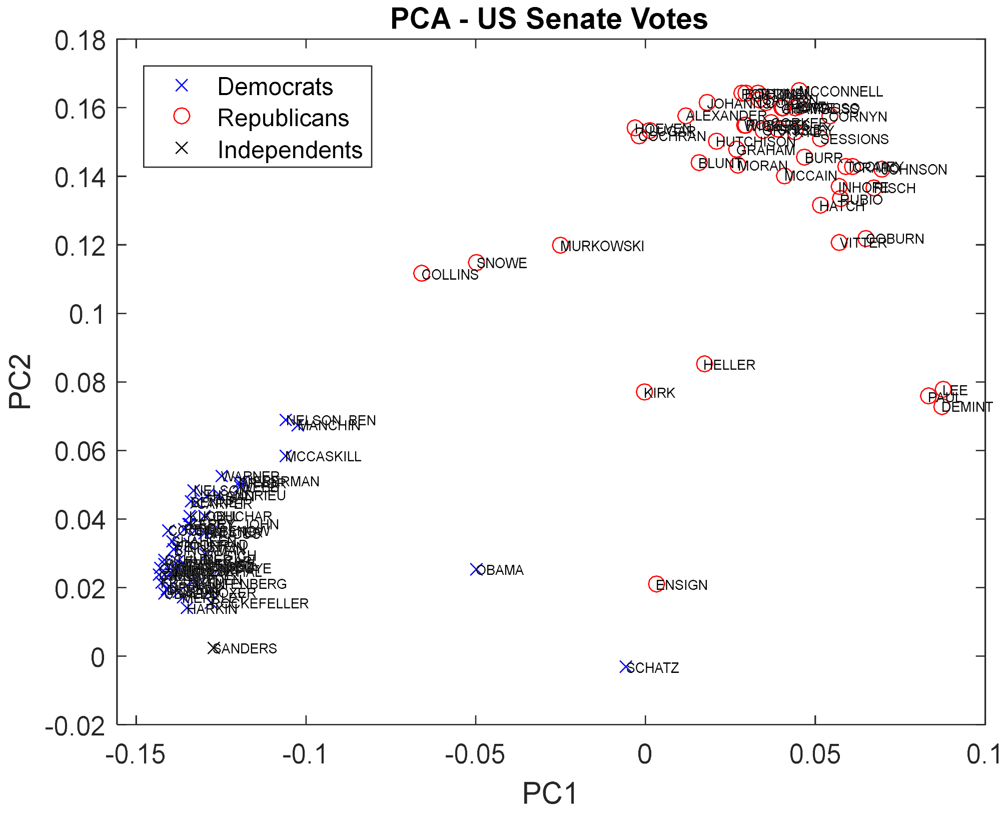
The SVD of the US Senate Voting Pattern was computed as required. The singular values were plotted. The plot above shows a number of dominant principle components between σ_1 and σ_5. The magnitude of the principle component decays sharply from σ_1 and σ_5. After σ_5 the principle components decay at an exponentially constant rate. This trend continues until $\sigma_{80,}$ where there is a gentle roll off.

(c) Create a scatter plot of the first and second columns of U. That is, each coordinate in the scatter plot will be ($u_1(j)$, $u_2(j)$). Color each coordinate by party affiliation (Democrat, Republican, or Independent).

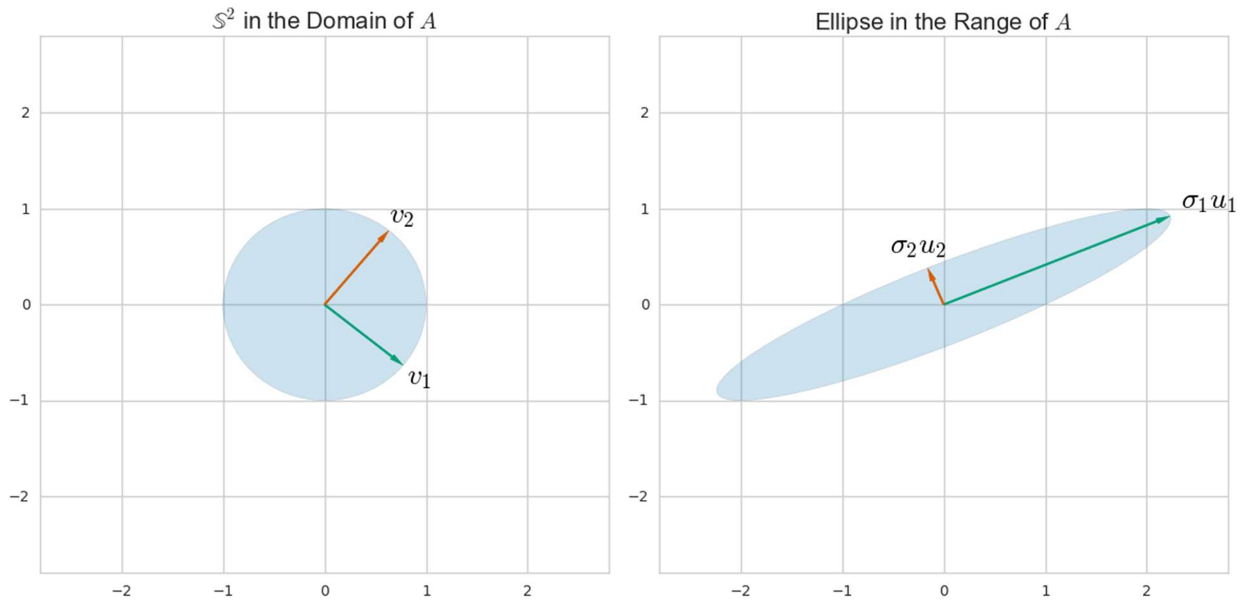



PCA - US House Votes
Democrats
Republicans
PC2
PC1
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
AMASH
PAUL
HELLER
WEINER
BARBER
WU
FILNER
BONAMICI
KUCINICH
JACKSON
INSLEE
GUTIERRE
HAHN
GRIJALVA
TOWNS
CARDOZA
RANGEL
SHULER
COSTELLO
SCHRADER
HOCHUL
PETERSON
MATHESON
BOREN
HOLDEN
COSTA
COOPER
DONNELLY
MCINTYRE
ALTMIRE
KISSELL
CHANDLER
BARROW
CRITZ
CUELLAR
SCHIFF
CONNOLLY
WALZ
CAMPBELL
BROUN
MCCLINTOCK
BACHMANN
MACK
GARRETT
AMODEI
JONES
AKIN
GIBSON
HANNA
FITZPATRI
CAPITO


PCA - US Senate Votes
Democrats
Republicans
Independents
PC2
PC1
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.15
-0.1
-0.05
0
0.05
0.1



(d) **Examining this plot, what do you think each coordinate represents? That is, what do you think u_1 and u_2 are capturing? Consider generating other plots involving u_1 and/or u_2 to aid in understanding. (Hint: Think about the transformation of a unit circle by the SVD.)**


$\mathbb{S}^2$ in the Domain of A
Ellipse in the Range of A
v_2
v_1
$\sigma_1 u_1$
$\sigma_2 u_2$
2
1
0
−1
−2
−2
−1
0
1
2

From the figure of a 2D transformation via SVD we can visualize a few useful definitions which hold for arbitrary dimensions.

- The lengths σ_i of the semi-axes of the ellipse are the *singular values* of A.
- The unit vectors u_i along the semi-axes of the ellipse are called the *"left" singular vectors* of A.
- The unit vectors v_i such that $Av_i = \sigma_i u_i$ are called the *"right" singular vectors* of A.

Because of the properties of SVD, the singular values will capture variance in decreasing order.

That is: $\sigma_1 > \sigma_2 > \ldots > \sigma_n$

Relating back to the problem this means that the first two columns of the u matrix, the vectors u_1 and u_2

represent the first two basis vectors of the new coordinate system, while the corresponding singular values, σ_1 and σ_2 represent their scalar magnitudes.

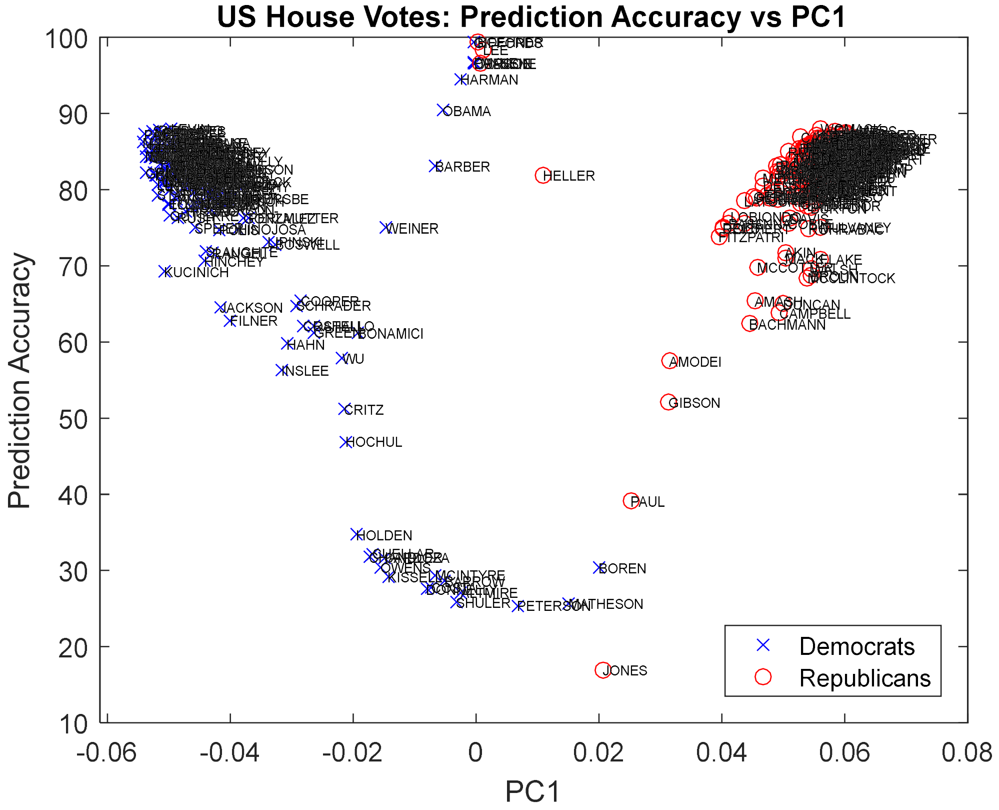
Singular Value Decomposition is a matrix factorization method utilized in many numerical applications of linear algebra such as principle component analysis. PCA allows one to reduce the dimensionality of a domain.

(e) Use a low rank approximation of the voting record using the first two dominant singular values 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 u1 vector. What do you observe? It can be instructive to plot the name of each representative on their respective point.

TABLE - LOW RANK APPROXIMATION OF US CONGRESS VOTING RECORD

United States House	Total Correct	Prediction Accuracy
House of Representatives	512,967	79.45 %
Senate	37,951	78.09 %

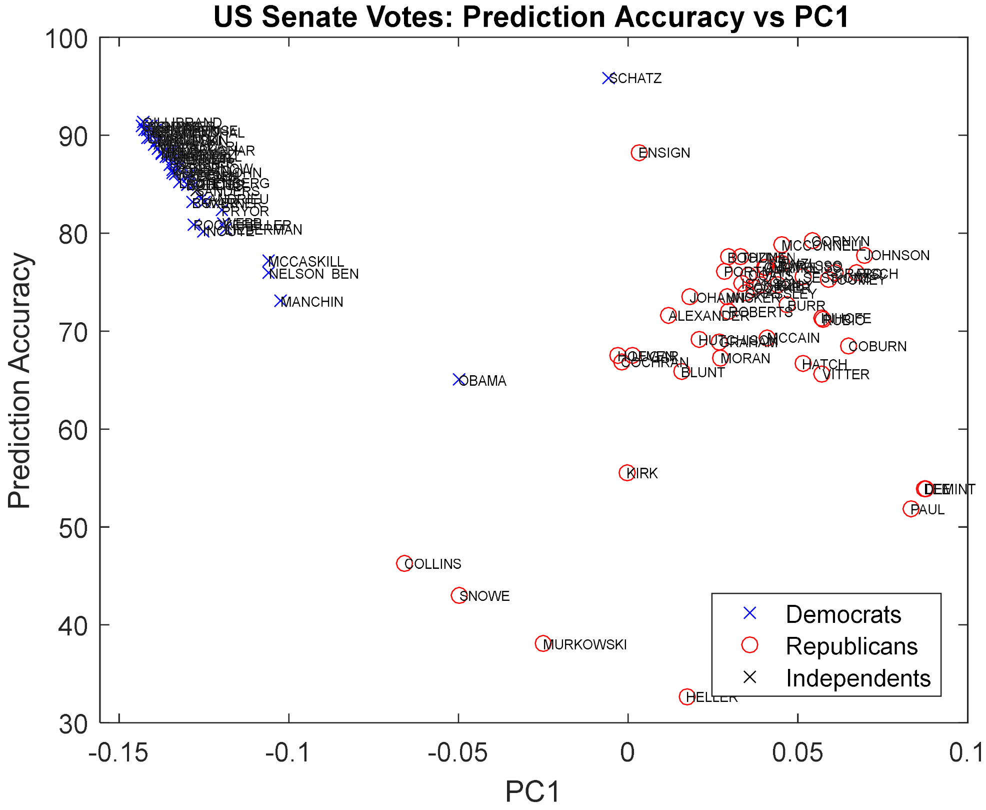
The rank 2 approximation of voting record is summarized above for both the House of Representatives and the Senate. The prediction accuracy of this approximation was 79.45 %, for the House of Representatives, and 78.09 %, for the Senate. This demonstrates how much variance the first 2 principle components (singular values) explain in the data. This prediction accuracy would be further enhanced by including more principle components (singular values): that is a higher rank approximation would provide even greater accuracy if required.



The plot above once again shows that PC1 is correlated to party affiliation.

- As PC1 decreases, it is more likely that the voter is a Democrat.
- As PC1 increases it is more likely that the voter is a Republican.
- There is generally less variance in predictive accuracy for voters as |PC1| increases
- There is generally more variance in predictive accuracy for voters as |PC1| decreases
- The relationship for predictive accuracy and PC1 appeared symmetric about PC1 = 0.01

Examining some of the voters with PC1 closer to the center, it appeared that these voters had a higher frequency of not voting the same way as the majority of their party. This was either via frequent abstention like President Obama, Senator Heller and Senator Lee, or via crossing the floor on issues, like Senator Jones.



The plot above once again shows that PC1 is correlated to party affiliation.

- As PC1 decreases, it is more likely that the voter is a Democrat.
- As PC1 increases it is more likely that the voter is a Republican.
- There is generally more variance in prediction accuracy for voters as |PC1| decreases
- The relationship although similar to the one seen US House, is not as symmetric.
- The grouping for the Republican senators has a larger spread, relative to the grouping of the Democratic senators.
  Examining some of the voters with PC1 closer to the center, it appeared that these voters had a higher frequency of not voting the same way as the majority of their party. This was either via frequent abstention like President Obama and Senator Heller, or via crossing the floor on issues, like Senator Murkowski.

(f) Repeat this exercise for the House data

Completed. See above.

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

TABLE - LOW RANK APPROXIMATION OF US CONGRESS VOTING RECORD

United States House	Prediction Accuracy K = 1	Prediction Accuracy K = 2	Prediction Accuracy K = 3
House of Representatives	64.69 %	79.45 %	81.01 %
Senate	54.54 %	78.09 %	79.92 %

The rank 2 matrix explained explains about 80 % of the voting variance. The PC1, the first singular value, appears to be highly correlated to voting along party lines. This means that 64.69 % and 54.54 % voting variance in the House of Representatives, and the Senate respectively is explained by PC1. PC1 values closer to the center indicate a tendency to vote against the majority of their party, either via frequent abstention like President Obama or via crossing the floor on issues, like Senator Jones. PC1 values closer to the edges correlate to frequently voting along party lines.