APM120, quiz #2, 2022: Solutions
Applied Linear Algebra and Big Data
Last updated: Monday 3rd April, 2023, 07:55

Read these instructions carefully: Please solve all problems, deriving, calculating, and showing explicitly all stages of your solution and explaining each step of each question. Do not use scratch paper; answers without intermediate steps will trigger a penalty. The number of points for each question is noted below, the total number of points is 110, and the final score is min(100, your points). Use a non-programmable calculator to convert your answers to decimal number format, carrying out calculations to **three significant digits**. Please box your final answers. Limit your essay responses to no more than the specified number of words, longer responses will be truncated. **Allowed aids:** only a simple calculator and TWO one-sided letter-sized pages with review material.

Start time: 6:30, end time including upload: 9:00.

Total time, including solving/ scanning via genius scan or a similar app/ uploading/tagging each question on gradescope (as in homework): **2.5 hours**, strictly enforced. Gradescope will terminate the session and submit uploaded material after 2.5 hours.

Enjoy!

1. (34 pts) Consider the linear set of equations $\mathsf{A}\mathbf{x} = \mathbf{b}$, and the SVD: $\mathsf{A} = \mathsf{U}\Sigma\mathsf{V}^T$,

```
A=[    1,     1,     1;
      -3,    -3,  -2.5];
b=[    2;
       4];
U=[-0.331,  0.944;
    0.944,  0.331];
Sigma=[ 5.22,     0,     0;
           0,  0.136,     0];
V=[-0.606, -0.365, -0.707;
   -0.606, -0.365,  0.707;
   -0.515,  0.857, -1.11e-16];
```

   (a) 50 words: **(i)** In general, if there are fewer equations than unknowns ($M < N$) and the rank of the matrix is equal to the number of equations ($r = M < N$), do you expect a solution exists? How many? Why? How would you define a useful and unique solution in such a case? **(ii)** What is a pseudo inverse, write the expression for it, explain how it is used to solve an underdetermined case.

   (b) Calculate the unique solution for the particular system given based on the recipe you outlined.

   (c) In general, no numerical calculations: **(i)** What is the condition number, and what is it used for? **(ii)** What is the effective rank, and how is it different from the rank.

   (d) **(i)** What is the rank of A? **(ii)** What is the effective rank of A if the noise in the RHS is 10%? Are we guaranteed to have a solution in that case? How would you solve for $\mathbf{x}$ then (explain with no calculations)? **(iii)** Calculate the condition number of A.

Solution:

   (a) (4) **(i)** if there are fewer equations than unknowns ($M < N$) and the rank of the matrix is equal to the number of equations ($r = M < N$), there will be an infinite number of solutions because $r = M$ guarantees that there cannot be any contradictory equations. A useful and unique solution is defined by requiring it to solve the equations and have the smallest norm possible.

   (4) **(ii)** $A^\dagger = V\Sigma^\dagger U^T$ where $\Sigma^\dagger$ has one over the singular values on the diagonal, and it is transposed relative to Σ. The solution is then $x = A^\dagger\mathbf{b}$.

   (b) (9) Solution:

```
Sigma_dagger=[ 0.192,     0;
                   0,  7.38;
                   0,     0];
A_pinv=[ -2.5,    -1;
```

```
          -2.5,    -1;
             6,     2];
x=[   -9;
      -9;
      20];
r=[    0;
       0];
```

(c) **(3) (i)** the condition number is calculated as σ_1/σ_r. It represents the maximum amplification of the error in the solution $|\delta x|/|x|$ relative to that in the rhs $|\delta b|/|b|$.

(3) (ii) Suppose the relative error in the matrix is G. The effective rank is then the p such that $\sigma_p/\sigma_1 > G$ but $\sigma_{p+1}/\sigma_1 < G$. Therefore, it is the number of singular values that are significantly different from zero based on the specified error level. The rank of a matrix is the number of nonzero singular values. It does not depend on the error level in the matrix elements but only on their values.

(d) **(4) (i)** the rank of A is 2.

(4) (ii) The effective rank of A if the noise in the RHS is 10% is 1. We are not guaranteed to have a solution in that case because $r < M$, and we can have equations that are contradictory within the level of accuracy. To solve for $\mathbf{x}$ then, we minimize residuals and require that the norm of the solution is as small as possible. The solution is again equal to A dagger times b.

(3) (iii) the condition number for A is

```
condition_number=[ 38.5];
```

2. (33 pts) Consider the data sets X, representing normalized monthly variations in the numbers of individuals infected in Cambridge by Covid (first line) and flu (second line) over $N = 5$ months. Similarly, the data set Y represents the same parameters, yet specifically within the Harvard community. Let the combined 4×5 data set be $\mathsf{F} = [\mathsf{X}; \mathsf{Y}] = \mathsf{U}\Sigma\mathsf{V}^T$, where,

```
X=[   1.68,   -0.92,   -2.32,   -0.12,    1.68;
      3.16,   -0.54,   -4.84,   -0.94,    3.16];
Y=[  -1.52,    0.28,    2.58,    0.18,   -1.52;
      1.68,   -0.52,   -2.22,   -0.62,    1.68];
U=[ -0.383,   0.878,   0.158,  -0.239;
    -0.753,  -0.444, -0.0819,  -0.478;
     0.379,    ????,  -0.695,  -0.598;
    -0.377,   0.124,  -0.697,   0.598];
diag_Sigma=[   8.85,   0.682,    ????,  1.85e-16];
V=[ -0.478,   0.124,  -0.237,  -0.699,    0.46;
      0.12,  -0.874,   0.146,  0.00815,   0.447;
     0.717,   0.248,  -0.474,  0.00815,   0.447;
      ????,   0.379,   0.801,  0.00815,   0.447;
    -0.478,   0.124,  -0.237,   0.715,   0.434];
F_reconstructed=[   1.62,    ????,   -2.43,  -0.405,    1.62;
                    ????,    -0.8,   -4.78,  -0.795,    3.19;
                   -1.61,   0.403,    2.41,     0.4,   -1.61;
                    1.59,    -0.4,   -2.39,  -0.398,    1.59];
```

(a) **(i)** Calculate the missing elements of U, V and Σ. **(ii)** 50 words: Interpret all elements of the first three columns of U, noting and explaining their signs and magnitudes. Take into account that small numbers may be interpreted as being effectively zero. Interpret the first V vector given your interpretation of the first U vector. *Note:* do not just indicate that there is a positive/negative correlation between this or that, but rather tell a brief(!) story of what this means. **(iii)** 50 words: What do the U SVD vectors represent in PCA in general? The V vectors? What can the singular values be used for in a PCA analysis, and how (give the formula)? **(iv)** 30 words: Explain why multivariate PCA may sometimes fail to extract the correct relation between two data sets. Why can MCA help? What information is lost in MCA?

(b) **(i)** Calculate the fraction of the total variance explained by each principal component. **(ii)** 100 words: Discuss: how do you decide how many U and V vectors are required to explain the data? What is the right choice in this case? How does the number of required PCs depend on the accuracy of the data? **(iii)** Calculate the missing elements in the $\mathsf{F}_{\text{reconstructed}}$ reconstructed from the first PC only. Compare the first and the third rows of the original data and the reconstructed one; interpret your findings in the context of the problem analyzed here rather than abstractly. *Hint:* calculate the ratios between the entries on the first and third rows.

Solution:

(a) **(6) (i)** missing elements of U, V and Σ:

```
U=[ -0.383,   0.878,   0.158,  -0.239;
    -0.753,  -0.444, -0.0819,  -0.478;
     0.379,   0.128,  -0.695,  -0.598;
    -0.377,   0.124,  -0.697,   0.598];
diag_Sigma=[   8.85,   0.682,   0.456,  1.85e-16];
V=[ -0.478,   0.124,  -0.237,  -0.699,    0.46;
      0.12,  -0.874,   0.146,  0.00815,   0.447;
     0.717,   0.248,  -0.474,  0.00815,   0.447;
     0.119,   0.379,   0.801,  0.00815,   0.447;
    -0.478,   0.124,  -0.237,   0.715,   0.434];
```

(6) (ii) Interpret the first three columns of U: The first one represents joint variability, with the first element of Y having the opposite trends from all three other elements. That is, the number of Covid Harvard cases goes down when Covid and flu numbers go up elsewhere. Perhaps because of the enhanced Harvard testing... Or because the virus is intimidated by the house deans? The second vector represents the variability in Cambridge only. Signs indicate that Covid and flu numbers in Cambridge that are not related to Harvard are anti-correlated. Third vector represents variability in Harvard only. And flu and Covid vary together.

Interpret the first V vectors: it shows an increase in time and then a decrease. This trend applies to the Harvard covid cases, but an opposite trend applies to the three others because they have negative signs in $\mathbf{u}_1$.

(6) (iii) U SVD vectors are the PCs. V vectors the nondimensional expansion coefficients/time series. The singular values are used to calculate the fraction of variance explained by mode i as $\sigma_i^2 / \sum_j \sigma_j^2$.

(5) (iv) Multivariate PCA may fail to extract the correct co-variability signal if the individual variability is larger than the joined one so that they occupy the first PC(s), and if the co-variability pattern is not orthogonal to the individual variabilities. MCA can help because it only extracts the co-variability, giving up on the independent variability modes in each data set.

(b) **(5) (i)** fraction of the total variance explained by each principal component:

```
fraction_variance=[  0.991;
                    0.00588;
                    0.00262;
                    4.31e-34];
```

(5) (ii) We decide how many U and V vectors are required based on their explained variance: only need to keep those that explain a significant part of the variance. In this case one mode already explains 99% which seems plenty for app purposes. The number of required PCs may decrease with increasing noise (reduced accuracy)

bec a PC that explains 5% may not be of interest if the noise level is 10%. **(iii)** Missing elements in $\mathsf{F}_{\text{reconstructed}}$:

```
F_reconstructed=[   1.62,  -0.407,   -2.43,  -0.405,    1.62;
                    3.19,    -0.8,   -4.78,  -0.795,    3.19;
                   -1.61,   0.403,    2.41,     0.4,   -1.61;
                    1.59,    -0.4,   -2.39,  -0.398,    1.59];
```

The ratios between the entries on the first and third rows indicate that the two vary together (the ratio is identical for all times). This makes sense as only a single time series vector is involved in the reconstruction. The trends in Harvard Covid numbers and the other three variables are opposite, as already deduced above.

3. (33 pts) Consider a data set X composed of 7 two-dimensional vectors and the corresponding Euclidean distance matrix given below.

   (a) Plot the data and calculate the missing elements of the distance matrix.

   (b) Use hierarchical clustering to cluster the data in X using the diameter cluster-merging criterion. Indicate each clustering by a circle in data space and number the circles. Use the distances given below. Use the data plot to guide your calculations. That is, calculate distances only where the plot does not clearly identify what the relevant smallest/ largest distances are, and discuss your assumptions explicitly at every step.

   (c) Plot the dendrogram based on the diameter after each merging.

   (d) Plot an elbow plot based on the diameter after each merging.

   (e) How is the appropriate number of clusters selected based on the dendrogram and elbow plot in the general case. What are possible challenges?

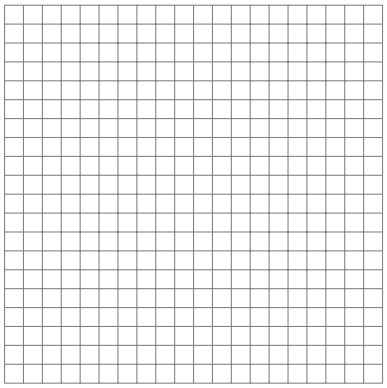
   (f) Calculate, for the cluster made of points 3,4,7, the center (clusteroid), radius, diameter, density (the number of points divided by radius), and variance.

   (g) 30 words: What is the cost of hierarchical clustering? Of k-means? Explain. When would k-means be *more* expensive than hierarchical clustering?

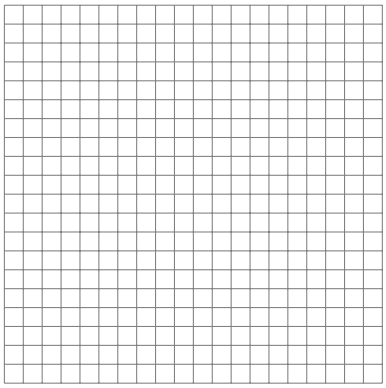
   (h) 20 words: draw an example 2d data set for which the 'single' cluster merging method is expected to fail, but where the clusteroid merging method should work reasonably well. Explain.

```
X=[ 8.5,   8.5,   0.5,   1.5,     8,   8.5,     2;
      8,   2.5,   5.5,     7,     3,   6.5,   5.5];
distances=[   0,   5.5,  ????,  7.07,  5.02,   1.5,  6.96;
            5.5,     0,  8.54,  8.32,  0.707,     4,  7.16;
           ????,  8.54,     0,   1.8,  7.91,  8.06,   1.5;
           7.07,  8.32,   1.8,     0,  7.63,  7.02,  1.58;
           5.02,  0.707,  7.91,  7.63,     0,  3.54,   6.5;
            1.5,     4,  8.06,  7.02,  3.54,     0,  6.58;
           6.96,  7.16,   1.5,  1.58,   6.5,  6.58,     0];
```

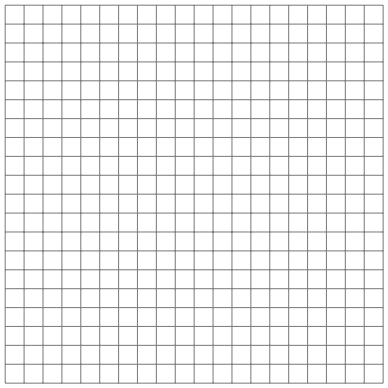
Data & clustering



dendrogram



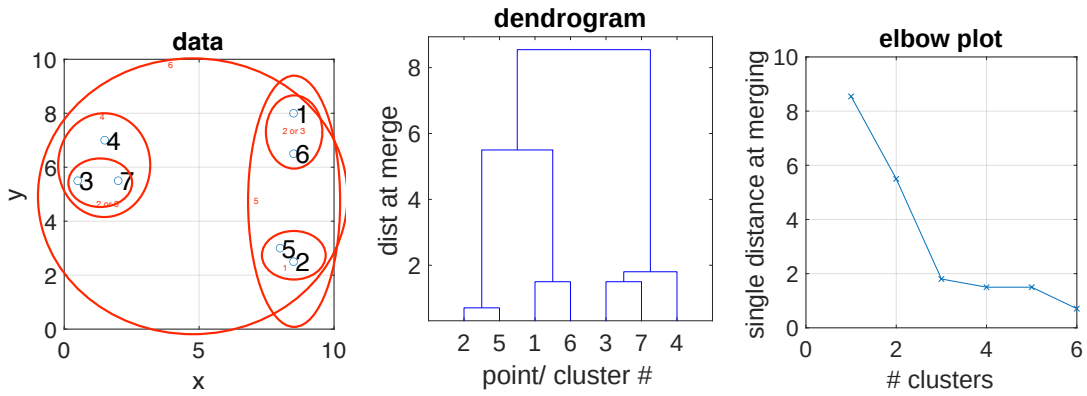
elbow plot



Solution:

(a) **(4)** Plot the data and calculate the missing elements of the distance matrix.

```
distances=[    0,   5.5,  8.38,  7.07,  5.02,   1.5,  6.96;
             5.5,     0,  8.54,  8.32, 0.707,     4,  7.16;
            8.38,  8.54,     0,   1.8,  7.91,  8.06,   1.5;
            7.07,  8.32,   1.8,     0,  7.63,  7.02,  1.58;
            5.02, 0.707,  7.91,  7.63,     0,  3.54,   6.5;
             1.5,     4,  8.06,  7.02,  3.54,     0,  6.58;
            6.96,  7.16,   1.5,  1.58,   6.5,  6.58,     0];
```



(b) **(5)** Hierarchical clustering: steps below are consistent with the dendrogram plot: a complete answer should show that the diameter of the merged cluster at each step is the smallest possible. In this simple example, this should be clear from the plot in most steps. But students should discuss this clearly as part of the solution.

```
1234567
1(25)3467
(16)(25)347
(16)(25)(37)4
(16)(25)((37)4)
((16)(25))((37)4)
(((16)(25))((37)4))
```

(c) **(4)** Dendrogram plot based on diameter (also known as the 'complete' distance): above.

(d) **(3)** Elbow plot based on diameter: above.

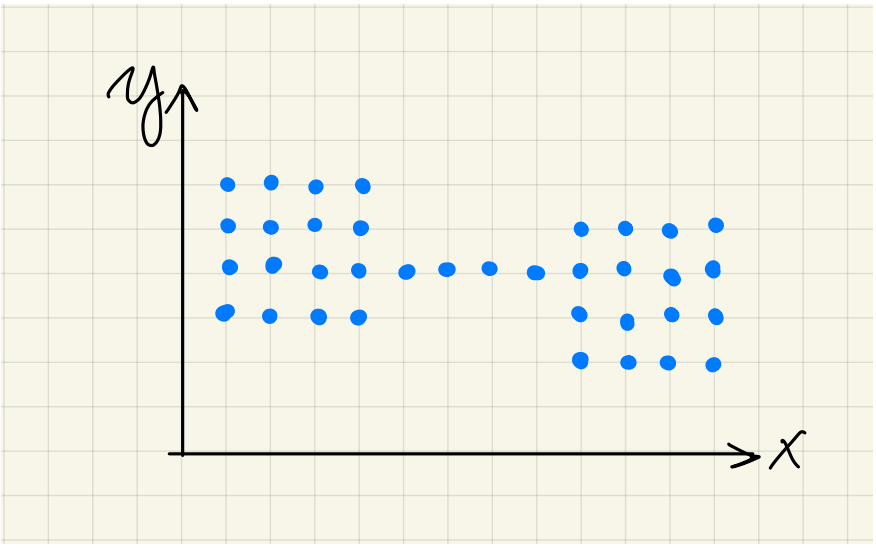
(e) **(4)** The appropriate number of clusters represents a sharp turn in the elbow plot (highest second derivative) and a large vertical distance between mergings shown in the dendrogram plot. Challenge is that sometimes data does not show a clear maximum in the second derivative/ or a clear elbow location. Similarly, the dendrogram does not show a clear location with a largest vertical separation between mergings.

(f) **(5)** For the cluster made of points 3,4,7:

```
center=[ 1.33;
            6];
distances_to_center=[ 0.972,  1.01,  0.833];
radius=[ 1.01];
diameter=[  1.8];
density=[ 2.96];
variance=[ 2.67]; (solution where variance was divided by 3 is
                   inconsistent with course notes but is acceptable)
```

(g) **(4)** The cost of hierarchical clustering is N^2 for calculating all distances between all points initially, plus various sorting and recalculating distances after each merging, which makes it even more expensive, $O(N^2 \log n)$. k-means cost is $N \times k \times \#iterations \times \#replicates$. k-means could be more expensive if $k \times \#iterations \times \#replicates > N$. That would typically mean that the data set is fairly small, so the cost does not matter anyway.

(h) **(4)** 20 words: an example 2d data set for which the 'single' cluster merging method is expected to fail, but clusteroid should work well:



Clusteroid method would work well as there are clearly two centers of data points here. Single would fail because the data points are regularly spaced, and there is no way to use single to tell where one cluster ends and the other begins.

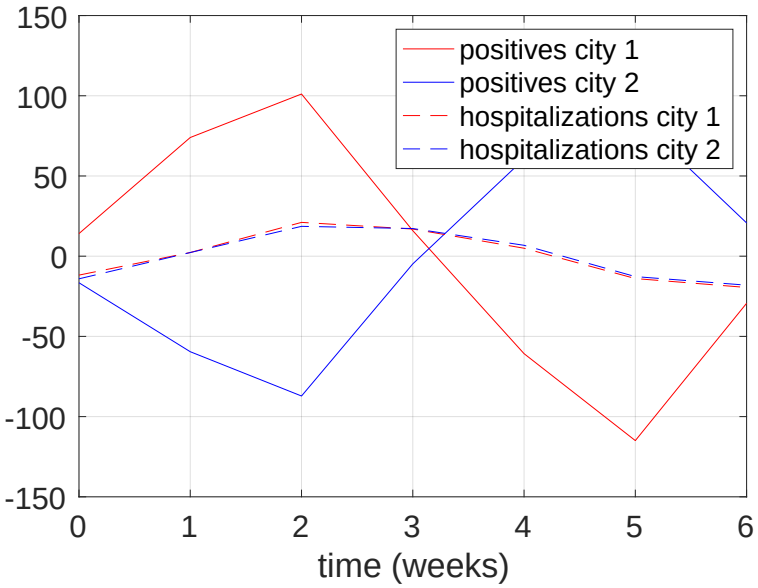
4. (10 pts) Consider the following two zero-mean data sets of the number of positive Covid tests in two cities over weeks zero to N (X) and hospitalization numbers for the same cities (Y). Plot the data and analyze the relationship between the two data sets, considering that one expects a delay between testing positive and being hospitalized. Interpret your results. *Hints:* **(1)** Find the optimal delay τ and perform the appropriate MCA of X during weeks 0 to $N-\tau$ vs. Y during weeks τ to N. **(2)** The SVD $\mathsf{A} = \mathsf{U\Sigma V}^T$ is given below.

```
X=[ 14.1,    74,   101,  15.9, -60.8,  -115, -29.4;
   -16.6, -59.5, -87.2,  -4.8,  60.6,  86.7,  20.7];
Y=[-11.8,  2.28,  21.1,  16.9,  4.95, -13.9, -19.5;
   -14.1,  2.27,  18.6,  17.2,  6.79, -12.7, -18.1];
A=[ 1078.08,  1018.02;
    -887.846,  -836.35];
U=[ -0.772281,  0.635282;
    0.635282,  0.772281];
Sigma=[ 1919.99,  0;
        0,  1.14351];
V=[ -0.727405,  -0.686209;
    -0.686209,  0.727405];
```

Note: This question is not asking you to repeat a calculation from HW or class but to think independently based on the material covered. You may make assumptions regarding whichever information you find needed yet unspecified.

Solution:

Plot data:



To find the optimal delay, one can either look at the correlation between the first line of X and the first line of Y as a function of delay (first element C_{11} of corresponding

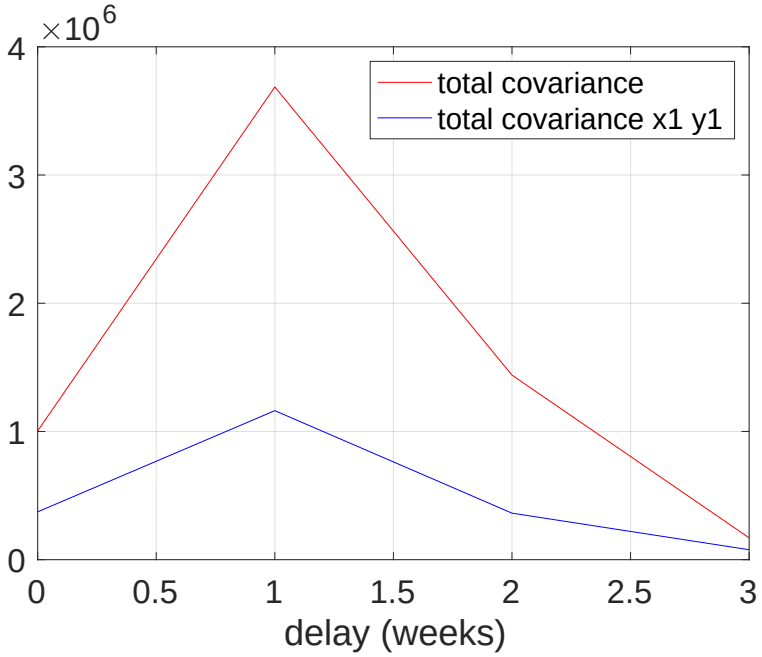
delayed covariance matrix),

$$\overline{x_1(t)y_1(t-\tau)},$$

or at the total covariance between the two data sets as a function of this delay. For a delay of one week, for example, these are calculated as,

```
C1=X(:,1:end-1)*Y(:,2:end)'/(N-1);
C1_1st_row=X(1,1:end-1)*Y(1,2:end)'/(N-1);
total_covariance(2)=sum(C1(:).^2);
total_covariance_1st_row(2)=C1_1st_row^2;
```

And one finds:



Or numerically,

```
total_covariance=[ 1.00452e+06,  3.68636e+06,  1.43942e+06,  170267];
total_covariance_x1_y1=[ 373192,  1.16225e+06,  362972,  77648.5];
```

In both cases, the optimal delay is one week. This indicates that Covid cases rise one week after infections do. Now analyze the appropriate delayed covariance C1 calculated above for one week, given by A in the question. Here are the expected elements in a complete solution/interpretation:

(a) **(2)** The optimal delay is 1 week, based on one of the two methods presented above.

(b) **(2)** The singular values indicate that one mode effectively explains all the covariance at a one week delay, so we only need that one.

(c) **(2)** The first U vector has a plus minus pattern, which means that Covid cases are anti-correlated in the two cities. The first V vector is also all negative, indicating that hospitalizations in both cities vary together and at a delay of 1 weeks after the Covid changes.

(d) **(2)** The fact that hospitalizations increase together in both cities while infections are anti-correlated suggests that each hospital serves both cities.

(e) **(2)** The amplitude of covid changes is much larger than of hospitalizations, indicating that only a small fraction of patients require hospitalizations.