APM120, quiz #2, 2019, **Solutions**
Applied Linear Algebra and Big Data
Last updated: Wednesday 6th April, 2022, 13:55

your name: _____, HUID: _____

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**. 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.

Start time: 7:00, end time: 9:00.

Enjoy!

1. (34 pts) Consider the linear set of equations $\mathsf{A}\mathbf{x} = \mathbf{b}$, and the SVD of the matrix A,

```
A=[    2,     1,     3;
       2,    -1,     1];
b=[    4;
       3];
[U,Sigma,V]=svd(A)
U=[-0.882, -0.472;
   -0.472,  0.882];
Sigma(1,1)=4.15;
Sigma(2,2)=1.67;
V=[-0.653,  0.491, -0.577;
   -0.0988, -0.81, -0.577;
   -0.751, -0.32,  0.577];
```

   (a) 50 words: (i) How many solutions do you expect exist? Explain. What assumption would you need to make to find a unique solution in this case? (ii) In general, if there are more equations than unknowns, how many solutions do you expect exist? How would you find a unique solution then?

   (b) 50 words: How is the pseudo inverse $\mathsf{A}^\dagger$ of a matrix $\mathsf{A}_{(M\times N)}$ defined (write the expression)? For an underdetermined system of equations $M < N$, what is the dimension of $\mathsf{A}\mathsf{A}^\dagger$? Of $\mathsf{A}^\dagger\mathsf{A}$? For $M < N$, do you expect each of these two products to be equal to the identity matrix? Under what condition?

   (c) Calculate the pseudo inverse of A and the unique solution for $\mathbf{x}$.

Solution:

   (a) (i) This is an underdetermined problem, with $r = M$, so an infinite number of solutions exist. A unique solution is obtained by requiring it to have the smallest possible norm. With $A = U\Sigma V^T$, this solution is given by $A^\dagger b$ where $A^\dagger = V\Sigma^\dagger U^T$, where sigma dagger is sigma transposed and with the inverse of the nonzero singular values along the diagonal. (ii) In this case there may be no solutions if the equations are contradictory, and we can find a single solution by requiring it to minimize the norm of the residuals $r = Ax - b$ instead of actually solving the equations.

   (b) The expression for the pseudo inverse $\mathsf{A}^\dagger$ is given above. If the rank of A satisfies $r = M < N$, then $(\mathsf{A}\mathsf{A}^\dagger)_{(M\times M)} = I$ and $(\mathsf{A}^\dagger\mathsf{A})_{(N\times N)} \neq I$.

   (c) Numerical solution, using,

$$\mathsf{A}^\dagger = V\Sigma^\dagger U^T$$
$$\mathbf{x} = \mathsf{A}^\dagger\mathbf{b}$$

   we find,

```
A=[ 2  1  3; 2  -1  1];
b=[ 4; 3];
[U,Sigma,V]=svd(A);
Sigma_dagger=Sigma';
for i=1:2; Sigma_dagger(i,i)=1/Sigma_dagger(i,i);end
A_pinv=V*Sigma_dagger*U';
A_pinv_matlab=pinv(A);
%% solve:
x=pinv(A)*b;
r=A*x-b;
AA_pinv=A*A_pinv;
A_pinvA=A_pinv*A;
condition_number=Sigma(1,1)/Sigma(2,2);
A=[    2,     1,     3;
       2,    -1,     1];
b=[    4;
       3];
U=[-0.882, -0.472;
   -0.472,  0.882];
Sigma=[ 4.15,     0,     0;
           0,  1.67,     0];
Sigma_dagger=[ 0.241,     0;
                   0,  0.599;
                   0,     0];
V=[-0.653,  0.491, -0.577;
   -0.0988, -0.81, -0.577;
   -0.751, -0.32,  0.577];
A_pinv=[-2.78e-17,  0.333;
         0.25, -0.417;
         0.25, -0.0833];
A_pinv_matlab=[-2.78e-17,  0.333;
                0.25, -0.417;
                0.25, -0.0833];
AA_pinv=[    1, -5.55e-17;
         -1.11e-16,     1];
A_pinvA=[ 0.667, -0.333,  0.333;
         -0.333,  0.667,  0.333;
          0.333,  0.333,  0.667];
x=[    1;
   -0.25;
    0.75];
r=[    0;
```

```
    1.78e-15];
cond=2.48421
```

2. (33 pts) Consider the data set `X` representing annual variations in the populations of mountain lions and bighorn sheep in Yosemite over 5 years, and the data set `Y` representing the variations in the annually averaged temperature and rainfall in the park.

```
X=[  1.6,  -0.4,   1.6,  -2.4,  -0.4;
     1.2,   0.2,   0.7,  -1.3,  -0.8];
Y=[-0.69,  0.05,  0.64,  0.21, -0.21;
    1.49, -1.25,  1.16, -2.41,  1.01];
```

Hints:

$$\mathsf{C} = \begin{bmatrix} -0.104, & 2.02 \\ ?, & ? \end{bmatrix}$$
$$\mathsf{U} = \begin{bmatrix} -0.907, & -0.42 \\ -0.42, & 0.907 \end{bmatrix}$$
$$\Sigma_{22} = 0.0426;$$

where $\mathsf{C} = \mathsf{X}\mathsf{Y}^T/5 = \mathsf{U}\Sigma\mathsf{V}^T$.

(a) (i) Calculate the covariance matrix and its SVD, using the above information and *hints*. 100 words for ii&iii: (ii) Interpret all elements of the covariance matrix, noting and explaining their signs and overall magnitude. Explain why the covariance elements make sense for the each of the specific variables in question in each data set. (iii) What do the U SVD vectors represent in Maximum Covariance Analysis in general? The V SVD vectors? What can the singular values be used for in a general MCA analysis and how (give the formula)?

(b) 100 words: What do the singular values indicate in this case (calculate...)? Discuss explicitly each of the 4 singular vectors and interpret them, and their importance, for this particular problem, in terms of what each of the `X` and `Y` variables represent.

(c) 40 words: Explain why multivariate PCA may sometime fail to extract the correct relation between two data sets.

Solution:

(a) (i) The covariance matrix and its SVD:

```
C=X*Y'/5;
[U,Sigma,V]=svd(C);
my_fprintf_array(C,'% 5.3g');
my_fprintf_array(U,'% 5.3g');
my_fprintf_array(Sigma,'% 5.3g');
my_fprintf_array(V,'% 5.3g');
C=[-0.104,  2.02;
```

```
   -0.095,  0.935];
U=[-0.907, -0.42;
   -0.42,  0.907];
Sigma=[ 2.23,     0;
           0,  0.0426];
V=[ 0.0601, -0.998;
   -0.998, -0.0601];
```

To complete the SVD of C by hand: calculate the eigenvalues of CC^T (simple, its a 2x2 matrix), take their square root, now you have Σ. Then use $C = U\Sigma V^T$ so that $C^T U = V\Sigma$ (equivalently, $C^T u_i = \sigma_i v_i$) to calculate V.

(ii) x1 and x2 uncorrelated with y1; x1 and, to a lesser degree, x2, positively correlated with y2. That is, lions and sheep are correlated with rain fall but not with temperature, because more rain means more grass and therefore more sheep, and that also means more lions which feed on the sheep. (iii) The U SVD vectors represent the X structures most correlated with the corresponding V vectors. The singular values provide the fraction of covariance explained by each pair of $\mathbf{u}_k$ and $\mathbf{v}_k$ vectors, given by $\sigma_k^2 / \sum_i \sigma_i^2$.

(b) In this case the fraction of variance explained by each of the two modes is,

```
fraction_variance=Sigma.^2/sum(Sigma.^2);
my_fprintf_array(fraction_variance,'% 5.3g');
fraction_variance=[    1;
                    1.32e-07];
```

so that effectively all of the covariance is explained by the first SVD mode. The first U vector includes both the lions and sheep with a negative sign, and the first V vector contains only the rainfall, again with a negative sign, indicating that sheep and rainfall are positively correlated, and lions and rainfall are positively correlated as well. The second U vector contains the sheep and lions with opposite signs, and the second V vector contains only the temperature. The relation between the two is not important, because the corresponding singular value is effectively zero.

(c) Multivariate PCA may fail because it would attempt to describe both the variance and the covariance of the two data sets. If the PC describing the co-variability is not orthogonal to the patterns describing the individual uncorrelated variability of the two data sets, the co-variability pattern will not be well-identified.

3. (33 pts) Consider the data set `X` composed of 7 two-dimensional vectors.

   (a) Plot the data and use hierarchical clustering to cluster the data in `X` using the 'single' distance measure (smallest distance between any two points in the two clusters). Indicate each clustering by a circle and number them. Use the distances given below.

   (b) Plot the dendrogram based on the clustering distance measure

   (c) Plot an elbow plot based on the clustering distance measure. Discuss the challenges in selecting an appropriate number of clusters based on the dendrogram and elbow plot.

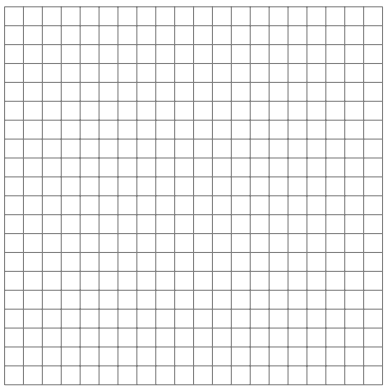
   (d) Calculate the center (clusteroid), radius, diameter, density (number of points divided by radius) and variance of a cluster made of points 2,6,7. If the definition of any of these quantities is not unique, select one sensible option and make it explicit in your solution.

   (e) 30 words: What is the cost of hierarchical clustering? Explain. Why is k-means less expensive?

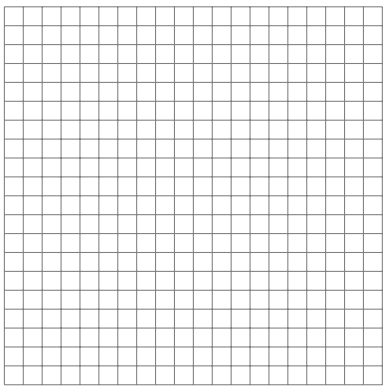
   (f) 30 words: Describe an example in which k-means is expected to fail when clustering using Euclidean distances, but where hierarchical clustering should work well. Explain.

```
X=[ 9,  7,  1,  8,  1,  5,  6;
    1,  7,  9,  1,  7,  5,  4];
distances=[    0,  6.32,  11.3,     1,    10,  5.66,  4.24;
            6.32,     0,  6.32,  6.08,     6,  2.83,  3.16;
            11.3,  6.32,     0,  10.6,     2,  5.66,  7.07;
               1,  6.08,  10.6,     0,  9.22,     5,  3.61;
              10,     6,     2,  9.22,     0,  4.47,  5.83;
            5.66,  2.83,  5.66,     5,  4.47,     0,  1.41;
            4.24,  3.16,  7.07,  3.61,  5.83,  1.41,     0];
```

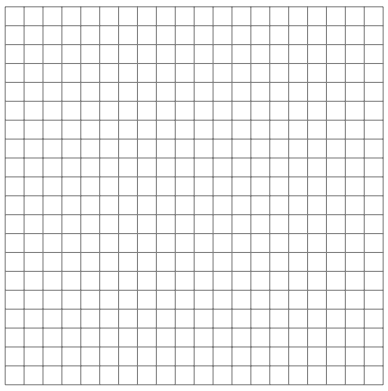
Data & clustering



dendrogram

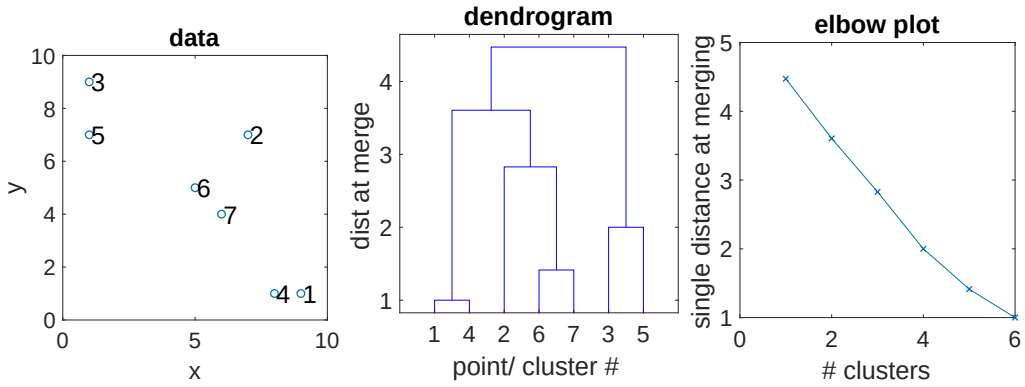


elbow plot



Solution:

(a) A plot of the data, clustering order (via the dendrogram), and elbow plot.



The implied clustering order from the dendrogram is,

```
1,2,3,4,5,6,7
(1,4),2,3,5,6,7
(1,4),2,3,5,(6,7)
(1,4),2,(3,5),(6,7)
(1,4),(3,5),(2,(6,7))
((1,4),(2,(6,7))),(3,5)
(((1,4),(2,(6,7))),(3,5))
```

(b) The elbow plot: above.

(c) The center, radius, diameter, density and variance of a cluster made of points 2,6,7:

```
% calculate:
XX=X(:,[2,6,7]);
center=mean(XX,2);
for i=1:3; distances_to_center(i)=norm(XX(:,i)-center); end
radius=max(distances_to_center);
distances_between_cluster_members = squareform(pdist(XX','euclid'));
diameter=max(distances_between_cluster_members(:));
density=3/radius;
variance=sum(distances_to_center.^2);
% print results:
my_fprintf_array(XX,'% 5.3g')
my_fprintf_array(center,'% 5.3g')
my_fprintf_array(distances_to_center,'% 5.3g')
my_fprintf_array(distances_between_cluster_members,'% 5.3g')
my_fprintf_array(radius,'% 5.3g')
my_fprintf_array(diameter,'% 5.3g')
```

```
my_fprintf_array(density,'% 5.3g')
my_fprintf_array(variance,'% 5.3g')
% results:
XX=[    7,     5,     6;
        7,     5,     4];
center=[    6;
         5.33];
distances_to_center=[ 1.94,  1.05,  1.33];
distances_between_cluster_members=[    0,  2.83,  3.16;
                                    2.83,     0,  1.41;
                                    3.16,  1.41,     0];
radius=[ 1.94];
diameter=[ 3.16];
density=[ 1.54];
variance=[ 6.66];
```

(d) Cost of hierarchical clustering is $O(N^3)$ because distances between all clusteroids need to be calculated at every step. Cost of k-means is less because the distances between all points need to be calculated only once for identifying the initial clusteroids, and then only the distance to all k clusteroids is required for each data point.

(e) k-means is expected to fail in the case of two concentric clusters, but Hierarchical clustering should work well using the 'single' linkage.

4. (10 pts) Consider the data set $\mathsf{X}(x,t)$ representing the sea surface temperature along the equator as function of east-west distance x and time t,

$$\mathsf{X}(x,t) = \sin(kx)\sin(\omega t), \quad x \in (0, L = 2\pi/k), \quad t \in (0, T = 2\pi/\omega).$$

Perform PCA on this data set by calculating the principal components, noting that sums are replaces with integrals, and vectors by functions in this case. You may run into an "integral equation" of the form $f(x) = \int_a^b K(x,y)g(y)dy$, try to solve it by differentiating both sides with respect to x.

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

Solution: the data is a continuous function, so the covariance matrix is given by,

$$\begin{aligned} C(x,y) &= \frac{1}{T}\int \mathsf{X}(x,t)\mathsf{X}(y,t)dt \\ &= \sin(kx)\sin(ky)\int sin^2(\omega t)dt \\ &= \frac{1}{2}\sin(kx)\sin(ky). \end{aligned}$$

To find the eigenvalues λ and eigenvectors (which are eigenfunctions $\phi(x)$ in this case), we write an integral equation that emulates $C\mathbf{v} = \lambda\mathbf{v}$,

$$\int C(x,y)\phi(y)dy = \lambda\phi(x).$$

substituting the covariance matrix we calculated,

$$\int \frac{1}{2}\sin(kx)\sin(ky)\phi(y)dy = \lambda\phi(x).$$

This is a Fredholm integral equation of the second kind, which is easily solved by taking the second derivative with respect to x of both sides,

$$-k^2\int \frac{1}{2}\sin(kx)\sin(ky)\phi(y)dy = \lambda\phi''(x).$$

Using the original integral equation,

$$-k^2\lambda\phi(x) = \lambda\phi''(x).$$

whose solution for the principal components (eigenfunctions) is $\phi(x) = \sin(kx)$. The time series is obtained by projecting the original data on the eigenfunctions,

$$T(t) = \int \mathsf{X}(x,t)\phi(x)dx = \int \sin(kx)\sin(\omega t)\ \sin(kx)dx \propto \sin(\omega t).$$

The original data can now be reconstructed using the principal component and the time series as,

$$\mathsf{X}_{\text{reconstructed}}(x,t) = \phi(x)T(t) = \sin(kx)\sin(\omega t).$$