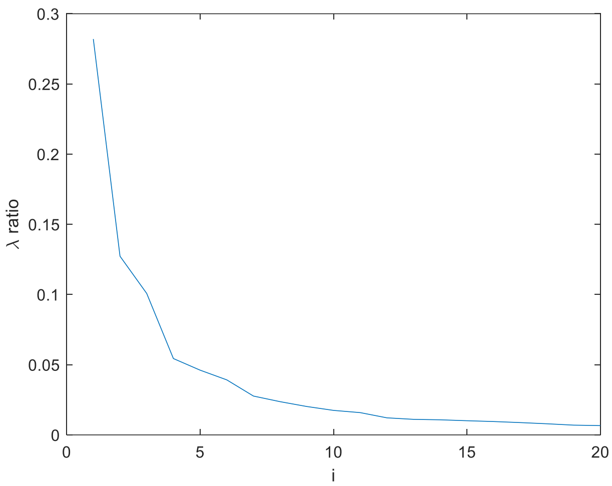
HW1

PENG, Han, 02/04/2023

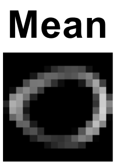
1. PCA experiments

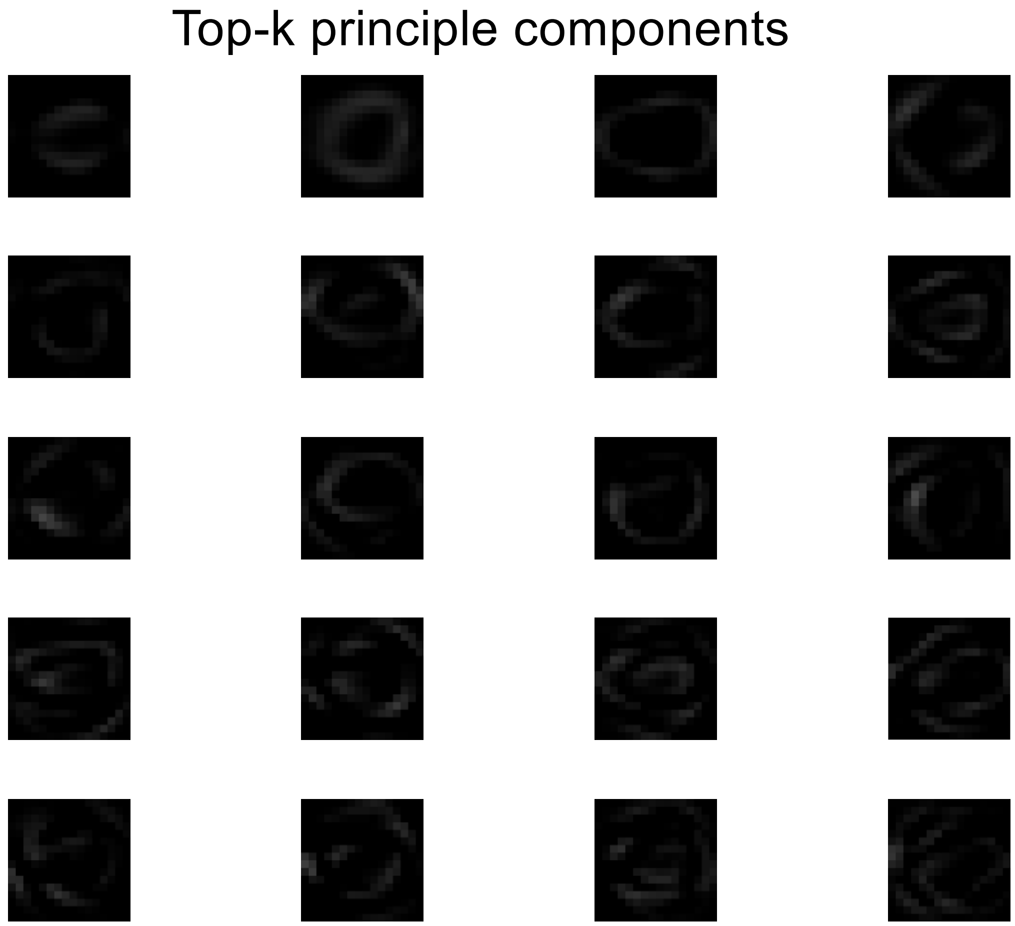
I selected 0 digit for this problem. MATLAB is used to perform computations and here we only present the visual results.

(d)

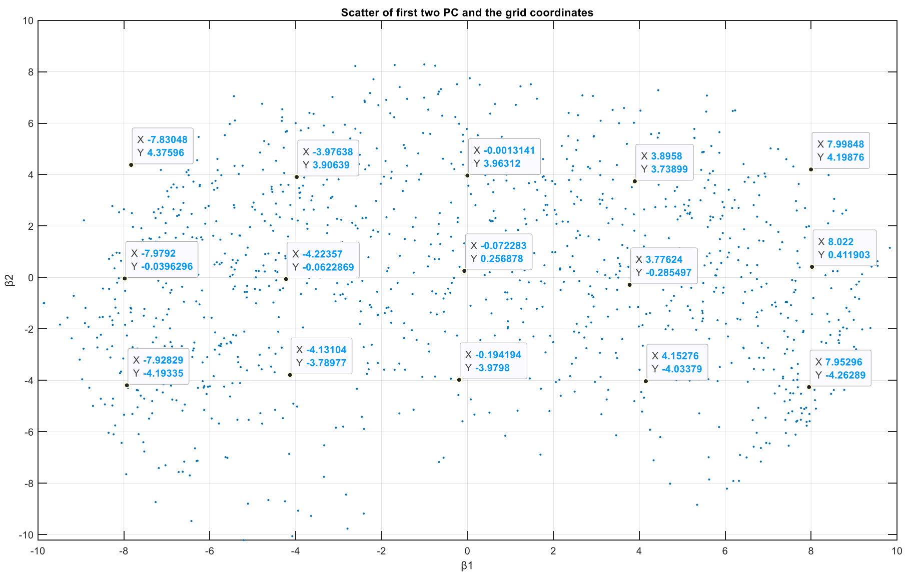


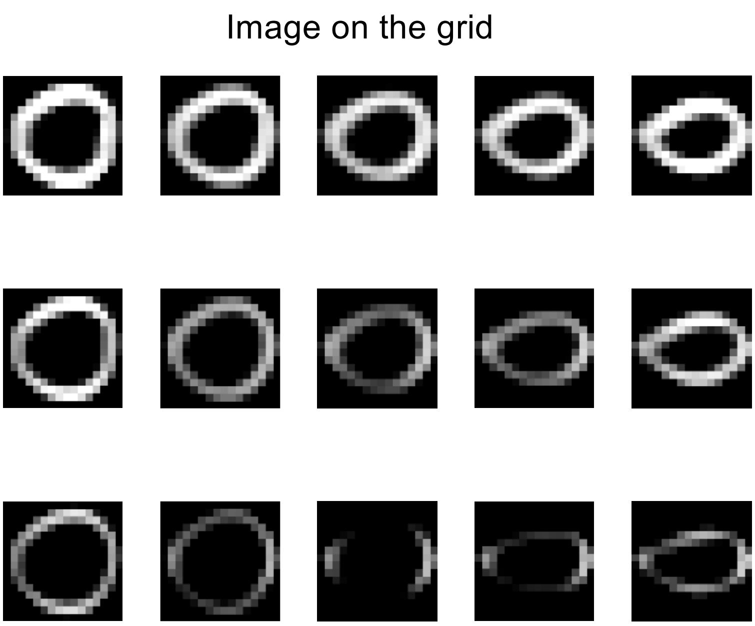
(e)




Top-k principle components

(g)


Scatter of first two PC and the grid coordinates
β1
β2
X -7.83048 Y 4.37596
X -3.97638 Y 3.90639
X -0.0013141 Y 3.96312
X 3.8958 Y 3.73899
X 7.99848 Y 4.19876
X -7.9792 Y -0.0396296
X -4.22357 Y -0.0622869
X -0.072283 Y 0.256878
X 3.77624 Y -0.285497
X 8.022 Y 0.411903
X -7.92829 Y -4.19335
X -4.13104 Y -3.78977
X -0.194194 Y -3.9798
X 4.15276 Y -4.03379
X 7.95296 Y -4.26289



2. MDS of cities

My first attempt is to select big cities in China and the distances between the cities are really large. However, the scatter plot of the cities' coordinates seems not correct and I realized that the distance between the cities given by the distancecalculator.net is based on the great circle distance, which is the arc length of two points in a sphere. On the other hand, the MDS algorithm I am using is based on Euclidean distance, which is suitable for planar setting. Therefore, I tried a second attempt using cities distance in the Bay area in the USA and this time the scatter plot looks much more reasonable.

(a)

The distance matrices of cities in China and Bay are shown below.

	Beijing	Shanghai	Hong kong	Changsha	Xi an	Kunming	Harbin	Wulumuqi
Beijing	0	1070.48	1973.7	1341.36	913.89	2087.97	1056.38	2412.66
Shanghai	1070.48	0	1227.83	883.46	1218.9	1958.64	1677.58	3269.78
Hong kong	1973.7	1227.83	0	669	1428.01	1208.93	2844.31	3418.39
Changsha	1341.36	883.46	669	0	777.19	1079.88	2291.52	2850.88
Xi an	913.89	1218.9	1428.01	777.19	0	1185.92	1970.27	2118.59
Kunming	2087.97	1958.64	1208.93	1079.88	1185.92	0	3138.97	2495.42
Harbin	1056.38	1677.58	2844.31	2291.52	1970.27	3138.97	0	3060.9
Wulumuqi	2412.66	3269.78	3418.39	2850.88	2118.59	2495.42	3060.9	0

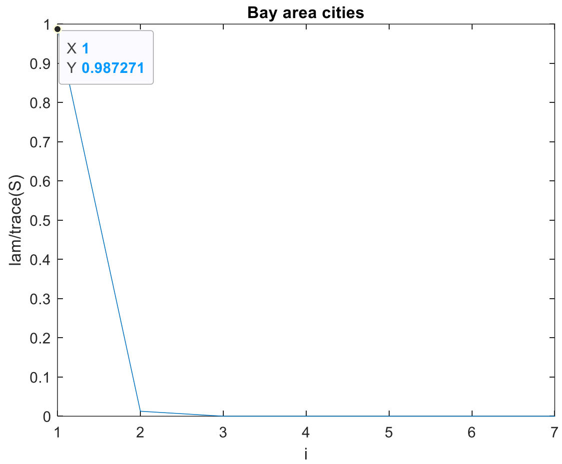
	berkeley	los angeles	san diego	sacramento	napa	san jose	livermore
berkeley	0	557.99	736.72	103.85	47.64	68.29	49.83
los angeles	557.99	0	179.25	581.28	595.42	491.41	512.46
san diego	736.72	179.25	0	759.97	774.57	669.76	691.55
sacramento	103.85	581.28	759.97	0	75.86	142.36	102.81
napa	47.64	595.42	774.57	75.86	0	112.67	83.28
san jose	68.29	491.41	669.76	142.36	112.67	0	39.61
livermore	49.83	512.46	691.55	102.81	83.28	39.61	0

(b)

I implemented my own MDS algorithm in MATLAB and I also compared it with the function provided by MATLAB. It should be noted that in for the MATLAB function one needs to provide a distance matrix while for my own implementation I need to provide a squared distance matrix.

(c)

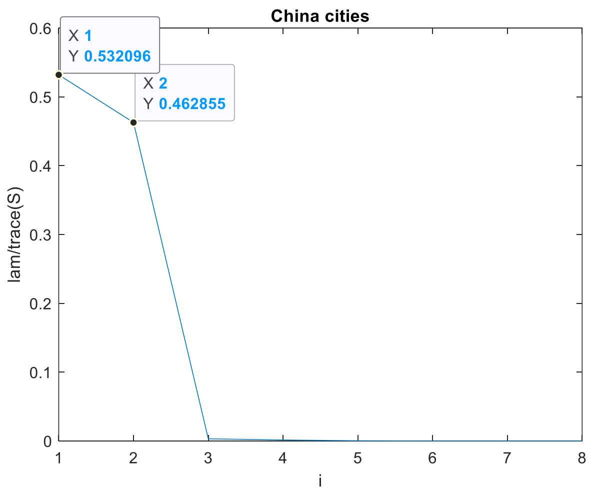
It is found that all the eigen values are positive. For Bay area cities, the first λ_1 occupies 98.7% of the sum of eigen values. For China cities, the first λ_1 and second λ_2 99.6% of the sum of eigen values. This suggests that there exists low-dimensional space that recovers most of the information of the data.



S

7x7 double

1	2	3	4	5	6	7
6.0155e+05	0	0	0	0	0	0
0	7.7451e+03	0	0	0	0	0
0	0	5.0706	0	0	0	0
0	0	0	4.1538	0	0	0
0	0	0	0	1.3738	0	0
0	0	0	0	0	0.0240	0
0	0	0	0	0	0	3.5831e-11



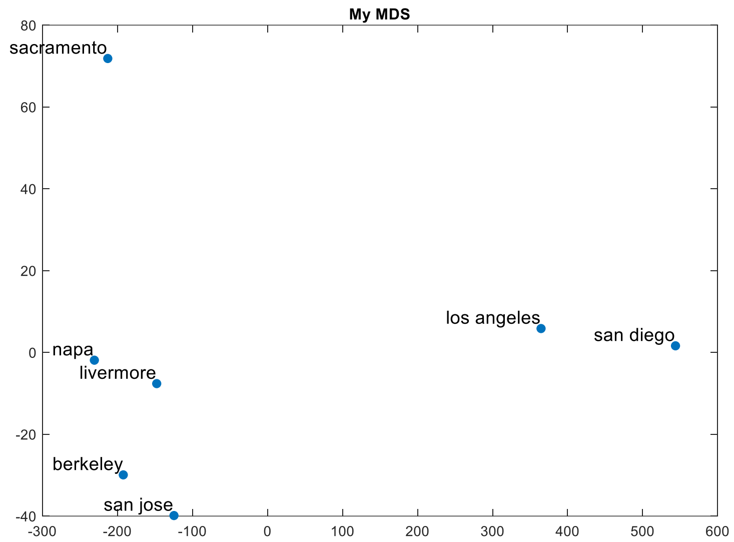
S

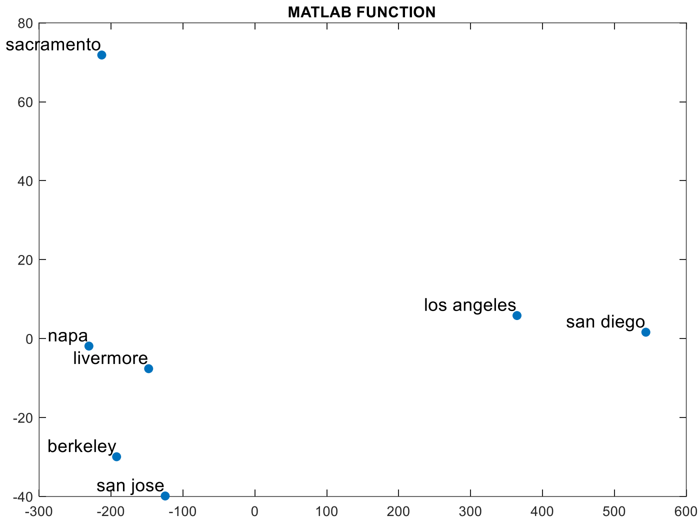
8x8 double

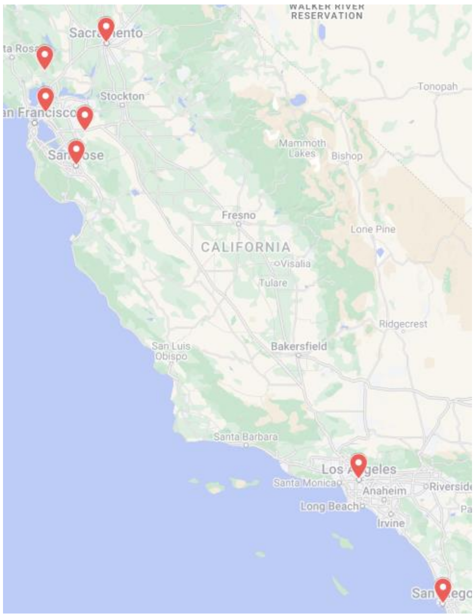
1	2	3	4	5	6	7	8
7.6491e+06	0	0	0	0	0	0	0
0	6.6537e+06	0	0	0	0	0	0
0	0	4.5865e+04	0	0	0	0	0
0	0	0	2.1533e+04	0	0	0	0
0	0	0	0	5.1791e+03	0	0	0
0	0	0	0	0	8.4671	0	0
0	0	0	0	0	0	4.1140	0
0	0	0	0	0	0	0	9.7945e-11

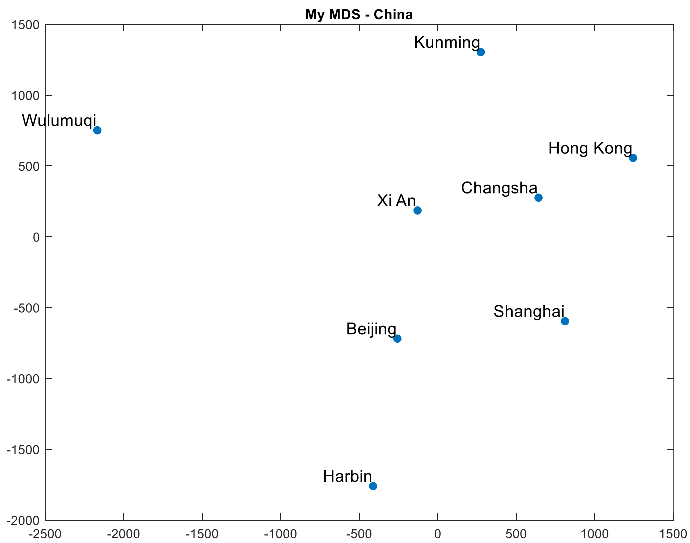
(d)

The scatter plots are shown below. We can see that my implementation matches with the MATLAB function. Furthermore, we notice that the coordinates computed by the MDS resemble the position of cities in the google map, demonstrating the effectiveness of the MDS algorithm. However, we also notice that the coordinates predicted by the MDS are rotated, as MDS does not preserve the origin and direction information. As mentioned above, the coordinates of China cities do not comply with the real map as the MDS is suitable for planar problem and will lose precision when the distance is large since real position of the cities is on a 3D sphere.




MATLAB FUNCTION
sacramento
los angeles
san diego
napa
livermore
berkeley
san jose
80
60
40
20
0
-20
-40
-300
-200
-100
0
100
200
300
400
500
600


WALKER RIVER RESERVATION
Stockton
Tonopah
Mammoth Lakes
Bishop
Fresno
CALIFORNIA
Visalia
Tulare
Lone Pine
Ridgecrest
San Luis Obispo
Bakersfield
Santa Barbara
Santa Monica
Anaheim
Long Beach
Irvine


My MDS - China
Kunming
Wulumuqi
Hong Kong
Changsha
Xi An
Shanghai
Beijing
Harbin
1500
1000
500
0
-500
-1000
-1500
-2000
-2500
-2000
-1500
-1000
-500
0
500
1000
1500

3. Positive Semi-definiteness

(a)

we can decompose x into the column space of the eigen vector matrix of K, which gives

$$x = \sum_{i=1}^{n} a_i v_i$$

Where a_i is the coordinate and v_i is the eigen vector.

$$Ax = A\sum_{i=1}^{n} a_i v_i = \sum_{i=1}^{n} a_i \lambda_i v_i$$

$$x^T Ax = A\sum_{i=1}^{n} a_i v_i = \left(\sum_{i=1}^{n} a_i v_i^T\right)\left(\sum_{i=1}^{n} a_i \lambda_i v_i\right)$$

Using the orthogonality of eigen vectors, we have

$$x^T Ax = \sum_{i=1}^{n} a_i^2 \lambda_i v_i^T v_i$$

Since $v_i^T v_i \geq 0$ and according to the condition $\lambda_i \geq 0$

Therefore $K \succcurlyeq 0$ if $\lambda_i \geq 0$.

On the contrary if there is a $\lambda_i < 0$, we can choose

$$x = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_j v_j + M v_j$$

Where M is a sufficient large number.

Then

$$x^T Ax = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_j^2 \lambda_j v_j^T v_j + M^2 \lambda_i v_i^T v_i < 0$$

Consequently, $K \succcurlyeq 0$ if and only if its eigenvalues are all nonnegative.

(b)

Since K is p.s.d, then there exists a matrix that satisfies $K = Y^T Y$ and $K_{i,j} = Y_i^T Y_j$

$$d_{i,j} = K_{i,i} + K_{j,j} - 2K_{i,j} = Y_i^T Y_i + Y_j^T Y_j - 2Y_i^T Y_j = \left(Y_i - Y_j\right)^T \left(Y_i - Y_j\right) = \left|\left|Y_i - Y_j\right|\right|^2$$

Therefore, $d_{i,j}$ is a squared distance function.

(c)

Denote $k = [K_{11}, K_{22}, \dots, K_{nn}]^T$, we can write matrix D as

$$D = k \cdot 1^T + 1 \cdot k^T - 2K$$

Now we want to prove an important equation:

$$-\frac{1}{2}H_\alpha D H_\alpha^T = H_\alpha K H_\alpha^T$$

Where $H_\alpha = I - 1 \cdot \alpha^T$, and $1^T\alpha = 1$

To see this, first we can write $-\frac{1}{2}H_\alpha D H_\alpha^T$ as

$$-\frac{1}{2}H_\alpha(k \cdot 1^T + 1 \cdot k^T - 2K)H_\alpha^T$$

Note that

$$1^T H_\alpha^T = 1^T(I - \alpha 1^T) = 1^T - 1^T\alpha 1^T = 0 \rightarrow H_\alpha k \cdot 1^T H_\alpha^T = 0$$

$$H_\alpha 1 = (I - 1 \cdot \alpha^T)1 = 1 - 1 \cdot \alpha^T 1 = 0 \;\rightarrow\; H_\alpha 1 \cdot k^T H_\alpha^T = 0$$

Therefore

$$-\frac{1}{2}H_\alpha D H_\alpha^T = H_\alpha K H_\alpha^T = B_\alpha$$

For any arbitrary x

$$x^T B_\alpha x = x^T H_\alpha K H_\alpha^T x = (H_\alpha^T x)^T K (H_\alpha^T x)$$

Since $K \succcurlyeq 0 \rightarrow (H_\alpha^T x)^T K (H_\alpha^T x) \geq 0 \rightarrow x^T B_\alpha x \geq 0$

Consequently

$$B_\alpha = -\frac{1}{2}H_\alpha D H_\alpha^T \succcurlyeq 0$$

(d)

we know that

$$x^T(A + B)x = x^T A x + x^T B x$$

Since $x^T A x \geq 0$ and $x^T B x \geq 0$. Therefore$x^T(A + B)x \geq 0$, which means

$$A + B \succcurlyeq 0$$

we can decompose x into the column space of the eigen vector matrix of A and B, respectively.

To prove the positive semi-definiteness of $A \circ B$, first we decompose A as

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T$$

$$A \circ B = \sum_{i=1}^{n} \lambda_i v_i v_i^T B = \sum_{i=1}^{n} \lambda_i diag(v_i) B diag(v_i)$$

Therefore

$$x^T(A \circ B)x = \sum_{i=1}^{n} \lambda_i x^T diag(v_i) B diag(v_i) x = \sum_{i=1}^{n} \lambda_i y_i^T B y_i$$

Where $y_i = diag(v_i)x$.

As A and B are both positive semi-definite, $\lambda_i \geq 0$ and $y_i^T B y_i \geq 0$. We have

$$x^T(A \circ B)x \geq 0 \rightarrow A \circ B \succcurlyeq 0$$

4. Distance

(a)

d^2 is not a distance function. Let's look at a one-dimensional problem $d_{xy} = |x - y|$.

Let $x_1 = 0, x_2 = 1, x_3 = 2$. $d_{12} = |x_2 - x_1| = 1$, $d_{23} = 1$, $d_{13} = 2$.

We can see that $d_{13}^2 = 4 > d_{12}^2 + d_{23}^2 = 2$, which violates the triangle inequality. Hence, d^2 is not a distance function.

(b)

$\sqrt{d}$ is a distance function. First it is easy to see $\sqrt{d}$ satisfies non-negative and symmetric property. Then we want to show that $\sqrt{d}$ satisfies triangle inequality.

First, we have

$$d(x,z) \leq d(x,y) + d(y,z)$$

Then

$$\sqrt{d(x,z)} \leq \sqrt{d(x,y) + d(y,z)} \leq \sqrt{d(x,y) + d(y,z) + 2d(x,y)d(y,z)} = \sqrt{d(x,y)} + \sqrt{d(y,z)}$$

Therefore $\sqrt{d}$ is a distance function.