Ans1

a) Support = 40% So 4 out of 10 transaction

Set 1

C_1

items	Support
Bread	7/10 = 70%
Coffee	4/10 = 40%
Sugar	3/10 = 30%
Milk	7/10 = 70%
Butter	4/10 = 40%
Cookies	2/10 = 20%
Eggs	3/10 = 30%

Support ≥ 40% ⇒

L_1

items	Support
Bread	70%
Coffee	40%
Milk	70%
Butter	40%

Set of 2

C_2

items	Support
Bread, Coffee	1 = 10%
Bread, Milk	4 = 40%
Bread, Butter	4 = 40%
Coffee, Milk	3 = 30%
Coffee, Butter	0 = 0%
Milk, Butter	2 = 20%

Support ≥ 40% ⇒

L_2

items	Support
Bread, Milk	40%
Bread, Butter	40%

C_3

items	Support
Bread, Milk, Butter	2 = 20%

No frequent itemset from C_3

b) All association Rules ↘

Association Rule	Support	Confidence	Confidence %	Valid
Bread → Milk	40%	4/7 = 0.57	57%	True
Milk → Bread	40%	4/7 = 0.57	57%	True
Bread → Butter	40%	4/7 = 0.57	57%	True
Butter → Bread	40%	4/4 = 1	100%	True

if Σ is diagonal then a perfect MVG. $V = (\text{Covariance Matrix})^{-1}$

$(x-\mu)^T \Sigma^{-1} (x-\mu) = K$ is Mahalanobis distance of vector $\vec{x}$ from mean value $\vec{\mu}$

>> when plotted = an ellipse or a p dimension ellipsoid in MVG

and this surface will be centred at μ.

$[x = [x_1, x_2]$ so $p = 2]$

Question 3

a)

iter	cluster assignments of data points (value A, B & C)								centroid locations		
	0.090	0.172	0.310	0.335	0.429	0.640	0.642	0.851	A	B	C
0	A	A	B	B	B	C	C	C	0.15	0.25	0.90
1	A	A	B	B	B	C	C	C	0.131	0.358	0.711
2	A	A	B	B	B	C	C	C	0.131	0.358	0.711

The algorithm converges after 2 iterations

b)

iter	cluster assignments of data points (value A, B & C)								centroid location		
	0.090	0.172	0.310	0.335	0.429	0.640	0.642	0.851	A	B	C
0	A	A	B	B	B	B	B	C	0.10	0.45	0.90
1	A	A	B	B	B	B	B	C	0.131	0.4712	0.851
2	A	A	B	B	B	B	B	C	0.131	0.4712	0.851

The algorithm converges after 2 iterations.

b)

c) Sum of Squared error for (A)

iter ① $SSE = (0.15 - 0.090)^2 + (0.15 - 0.172)^2 + (0.25 - 0.310)^2 + (0.25 - 0.429)^2 + (0.25 - 0.335)^2 + (0.90 - 0.640)^2 + (0.90 - 0.642)^2 + (0.90 - 0.851)^2$

$= 0.00036 + 0.000484 + 0.0036 + 0.032 + 0.067 + 0.066 + 0.0072 + 0.002$

$= 0.1818$

iter ② $SSE = (0.131 - 0.090)^2 + (0.131 - 0.172)^2 + (0.358 - 0.310)^2 + (0.358 - 0.429)^2 + (0.358 - 0.335)^2 + (0.711 - 0.640)^2 + (0.711 - 0.642)^2 + (0.711 - 0.851)^2$

$= 0.0016 + 0.0016 + 0.0023 + 0.0050 + 0.0050 + 0.0047 + 0.0196 + 0.0005$

$=$

iter 3 SSE = 0.0403 (As algorithm converges)

d) Sum of squared error for (b)

iter ① $SSE = (0.10 - 0.090)^2 + (0.10 - 0.172)^2 + (0.45 - 0.310)^2 + (0.45 - 0.429)^2 + (0.95 - 0.640)^2 + (0.45 - 0.642)^2 + (0.90 - 0.851)^2 + (0.45 - 0.335)^2$

$= 0.0001 + 0.0051 + 0.0196 + 0.0004 + 0.0361 + 0.0368 + 0.0024 + 0.0132$

$= 0.1137$

iter ② $SSE = (0.131 - 0.090)^2 + (0.131 - 0.172)^2 + (0.4712 - 0.310)^2 + (0.4712 - 0.429)^2 + (0.4712 - 0.335)^2 + (0.4712 - 0.640)^2 + (0.4712 - 0.642)^2 + (0.851 - 0.851)^2$

$= 0.0016 + 0.0016 + 0.0259 + 0.0018 + 0.0289 + 0.0284 + 0 + 0.0185$

$= 0.1068$

iter ③

SSE = 0.1068 as algorithm converges

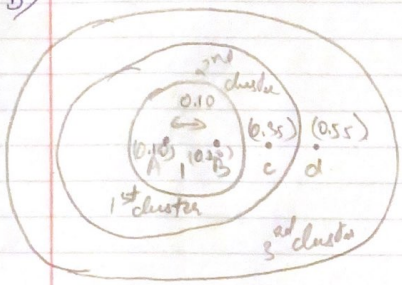
⇒ Solution (A) with initial centroids as (0.15, 0.25, 0.90) gives a better solution in terms of SSE

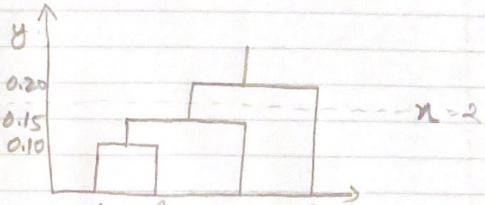
4) Points → Point (A) (B) (C) (D)
[0.10, 0.20, 0.35, 0.55]

a) Pair wise Euclidian distance →

	A	B	C	D
A	0	0.10	0.25	0.45
B	0.10	0	0.15	0.35
C	0.25	0.15	0	0.20
D	0.45	0.35	0.20	0

b)



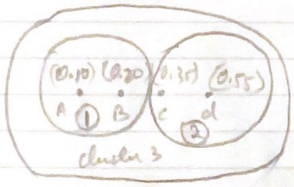


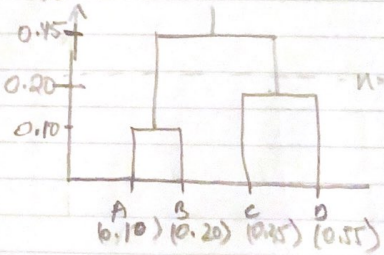
$SSE_{n=2} = (0.216 - 0.10)^2 + (0.716 - 0.20)^2 + (0.716 - 0.35)^2 + (0.55 - 0.55)^2$

$= 0.0134 + 0.0002 + 0.0149$

$= \boxed{0.0315}$

c)





$SSE_{n=2} = (0.15 - 0.10)^2 + (0.15 - 0.20)^2 + (0.45 - 0.35)^2 + (0.4 - 0.5)^2$

$= 0.0025 + 0.0025 + 0.9 + 0.01 \Rightarrow \boxed{0.025}$

d)

iter	
0	
1	
2	

$SSE_{n=2} = ($

The Mase

K-means

5)

Algorithm A
cluster1
cluster2

① $\Sigma_{i,j_A} \binom{n_{ij}}{2} =$

$\Sigma_{i_A} \binom{a_i}{2} =$

$\Sigma_{j_A} \binom{b_j}{2} =$

$ARI_A = 1$

$\frac{1}{2}[2475 +$

$ARI_A =$

Algorithm

2462.5 - 1237.5 1240 4950 1225 380

280 3.03 -16.96 2762.5 -

d)

iter	cluster assignment of data points (A, B, C, & D) A(0.10)	B(0.20)	C(0.35)	D(0.55)	centroid location C_1	C_2
0	C_1	C_1	C_2	C_2	0.15	0.5
1	C_1	C_1	C_2	C_2	0.15	0.45
2	C_1	C_1	C_2	C_2	0.15	0.45

$SSE_{k=2} = (0.15-0.10)^2 + (0.15-0.20)^2 + (0.45-0.35)^2 + (0.45-0.55)^2$
$= 0.025$

The Mase Method in (c) produces similar results and Max method and K-means with Euclidian distance produce a solution with lowest SSE

e)

Algorithm A	ground truth Class 1	Class 2
cluster 1	10	35
cluster 2	40	15

Algorithm B	ground truth Class 1	Class 2
cluster 1	35	40
cluster 2	15	10

f) $\sum_{ij}^{n_{ij}} \binom{n_{ij}}{2} = \binom{10}{2} + \binom{35}{2} + \binom{40}{2} + \binom{15}{2}$
$= 45 + 595 + 780 + 105 = 1525$

$\sum_i^{a_i} \binom{a_i}{2} = \binom{45}{2} + \binom{55}{2}$
$= 990 + 1485 = 2475$

$\sum_j^{b_j} \binom{b_j}{2} = \binom{50}{2} + \binom{50}{2}$
$= 1225 + 1225 = 2450$

$ARI_A = \frac{1525 - [2475 \times 2450]/\binom{100}{2}}{\frac{1}{2}[2475 + 2450] - [2475 \times 2450]/\binom{100}{2}}$

$ARI_A = 0.2424$

$\sum_{ij}^{n_{ij}} \binom{n_{ij}}{2} = \binom{35}{2} + \binom{40}{2} + \binom{15}{2} + \binom{10}{2}$
$= 1525$

$\sum_i^{a_i} \binom{a_i}{2} = \binom{75}{2} + \binom{25}{2}$
$= 2775 + 300 = 3075$

$\sum_j^{b_j} \binom{b_j}{2} = \binom{50}{2} + \binom{50}{2}$
$= 1225 + 1225 = 2450$

$ARI_B = \frac{1525 - [3075 \times 2450]/\binom{100}{2}}{\frac{1}{2}[3075 + 2450] - [3075 \times 2450]/\binom{100}{2}}$

$ARI_B = 0.0024$

Algorithm A is better in terms of ARI