Q1

*K* = 6

N = 240 x 180 = 43200

24 x 6 + 43200 x log26 = 111814 bits

compression ratio: 111814 / 1036800 = 10.8%

Q2

The error curve descends at the beginning and the algorithm will converge after some M steps, so the final EM cycle produces no changes in either the assignments or the prototype vectors, the error curve becomes flat.

The new curve probably not start and/or finish at the same points. Because the initial cluster centres are randomly selected in each experiment, the selection of initial cluster centres will affect the result of clustering.

Q3

Manhattan distances:

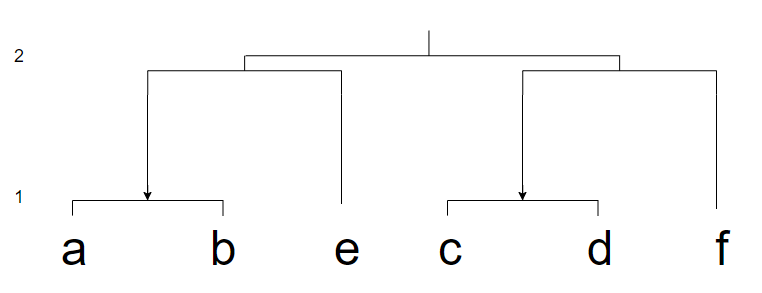
a – b : 1 b – c : 6 c – d : 1 d – e : 4 e – f : 5

a – c : 2 b – d : 7 c – e : 3 d – f : 3

a – d : 6 b – e : 3 c – f : 2

a – e : 2 b – f : 8

a – f : 7



Q4

Row 1: Circular distribution dataset, non-Euclidean distances, uneven cluster sizes, variable cluster density.

Row 3: There is no obvious boundary between different classes of data distribution, many noises, possibly

connectivity constraints.

Row 5: Gaussian mixtures model, the data is normally distributed by class, has clear boundaries, flat

geometry, good for density estimation.

Q5

Use the RANDOM technique.

RANDOM technique outperformed FA technique, RANDOM technique had a smaller standard deviation and a wider square-error values (x axis) range for hit counts, induced a more effective behaviour to the K-Means algorithm and made the K-Means algorithm exhibit a more robust behaviour, decreased the sensitivity of the K-Means algorithm to initial starting conditions.