Solutions to exercises: week 5

Exercise 5.1
(a) At the finest level, there are 14 clusters. These small clusters are organised in three larger clusters.
(b) This is an artificial problem. However, the desired number of clusters will usually depend on the application. If the small clusters have a physical meaning, such as e.g. individual species while the three larger clusters represent animal kingdoms, then the specific problem which the biologist is studying will determine the appropriate number of clusters, i.e. the level of detail considered.

Exercise 5.2
(a) There are no vertical stems that are distinctly longer than the other. The tree grows gradually, not in leaps and bounds.
(c) In terms of the lengths of the stems there is no difference.

Exercise 5.3
(a) These lengths correspond to the distances (single, average or complete linkage) between the two clusters being joined by the bridge consisting of two stems and a vertical bar.
(b) The closer together the samples in a cluster and the further apart the clusters, the better the clustering.
(c) Long stems correspond to large distances between clusters and therefore a potential point to cut the dendrogram.
(d) Yes, in the single linkage dendrogram, there are basically two lengths of vertical stems: those indicating the joining of the smallest clusters, and those indicating the joining of the larger three clusters. This is less pronounced with complete linkage.
(e) The average linkage dendrogram is roughly similar to the complete linkage dendrogram.

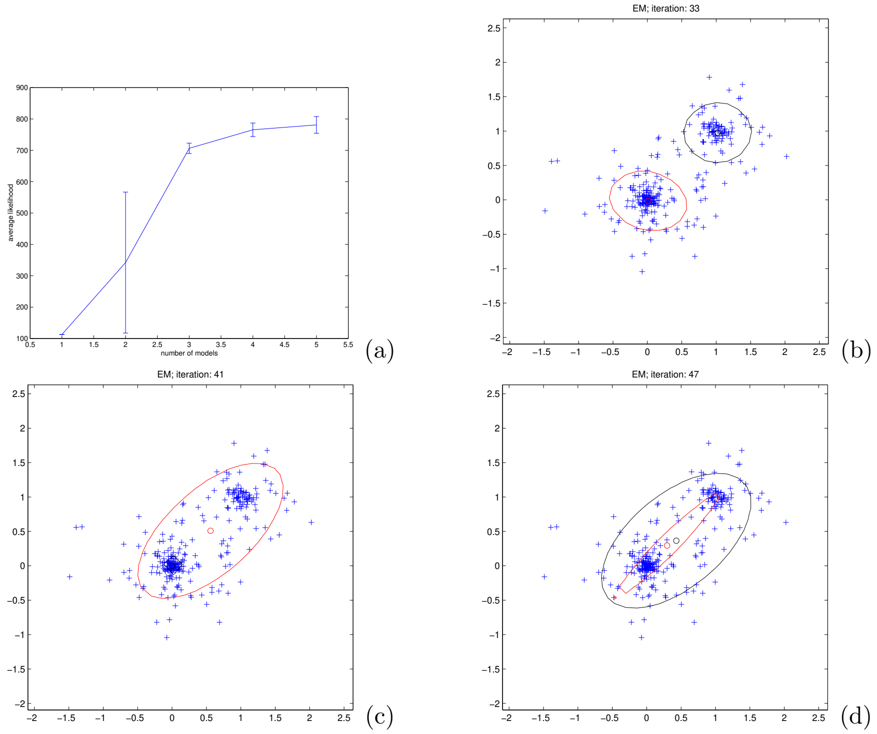
Exercise 5.5
(a) When samples become prototypes, the prototype is always amongst the samples. This obviously reduces the number of possible initial conditions, but also reduces the possibility of a situation where one prototype ends up without any samples.

Exercise 5.6
(a) K-means is better suited to datasets with spherical clusters, and should therefore work better on the `messy`dataset.

Exercise 5.7
(b) The algorithm has converged, i.e. no alteration of the prototype positions results in a decrease of the within-scatter, but the within-scatter of this solution is significantly higher than the within-scatter associated with solutions where there is a single prototype per cluster. In order to reach the global minimum, where each prototype is in a single cluster, one of the prototypes now sharing a cluster with another prototype needs to move towards the cluster without a prototype. In order to achieve this, the distances between the objects already assigned to the prototype that needs to move will increase, which will result in an *initial* increase in the within-scatter. Since the K-means algorithm only decreases the within-scatter, it remains stuck in this local minimum.
(c) It does not suffer from this problem, since it is a deterministic algorithm, independent of the random initial conditions.
(d) Run the algorithm several times with different initial settings, and choose the best result.

Exercise 5.8
(a) `'circular'`: diagonal covariance matrix with *equal* variances on the diagonal
`'aligned'`: diagonal covariance matrix with *unequal* variances on the diagonal
`'gauss'`: full covariance matrix, i.e. no constraints on the entries.
(b) The optimal k should be 2.
(c) The log-likelihood vs. the number of models is depicted in figure (a), the desired solution with two models in (b) and two frequently occurring undesirable solutions with two

models in figures (c) and (d). These undesirable solutions are the reason why the variance is so large at two clusters and why the curve does not flatten at two, but at three clusters.



Exercise 5.9

(a) A large jump in the fusion graph implies that two clusters were merged that were, relative to the other cluster distances, quite far apart.

Exercise 5.10

(a) At $g = 3$.

(b) The fusion graph is ambiguous about the exact number of clusters: either two or three clusters are associated with large fusion jumps of roughly the same magnitude. This is due to the complete linkage distance: distance between furthest samples of clusters at $(0, 0)$ and $(1, 1)$ (which are merged to result from a total of two clusters) is roughly half the distance between the distance between the furthest samples in the cluster at $(0, 0)$ and $(2, 2)$, which are merged to result in a single cluster. In single linkage this does not occur, since the minimal distances between the clusters at $(0, 0)$ and $(1, 1)$ is the same as the minimal distance between the clusters at $(1, 1)$ and $(2, 2)$.

Exercise 5.11

(a) There are two pronounced jumps, at 3 and 14 clusters.

Exercise 5.12

(a) At three clusters, since the largest fusion jump occurs at $g = 3$.

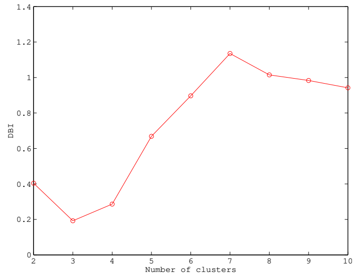
(b) No, three outliers constitute the members of two of the three clusters with all the other samples in the third cluster.

(c) While the size of the fusion jump is not very convincing (not much larger than the other jumps) the resulting clustering is better. This stems from the fact that complete linkage is less prone to outliers than single linkage.

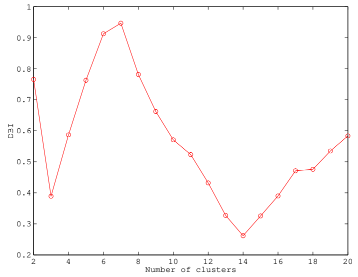
Exercise 5.13

(b) It should have a minimum at three clusters.

(c) The DBI curve is shown in the figure below, with a clear minimum at three clusters.



(d) There is a pronounced minimum at 3 clusters and a slightly larger minimum at 16 clusters. The first minimum (3 clusters) corresponds to the situation where the smaller clusters are grouped in three large clusters (four in top-left, four in top-right and six at the bottom) while the minimum at 16 corresponds to the fine cluster structure in the data. The peak at sixteen is more pronounced since the ratio of the maximal within-scatter to the minimal distance between any pair of these clusters is smaller than for the three cluster configuration.


DBI
Number of clusters