Multivariate Analysis (slides 10)

- Today we present some extensions of k-means clustering.
- We will also provide discussion on 'model-based clustering'.
- Previously we considered hierarchical clustering and k-means clustering, which are nonparametric (distribution free) in nature.
- Model-based clustering makes use of statistical models and is thus parametric in nature.

Extensions of k-Means Clustering

- Several possible extensions to k-means clustering have been proposed.
- First, recall that k-means clustering assigns observations to the group which has the closest centre in terms of squared Euclidean distance.
- That is, an observation $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{im})$ is assigned to group k so that $d(\mathbf{x}, \mu_k)$ is minimized, where

$$d(\mathbf{x}_i, \mu_k) = \sum_{j=1}^{m} (x_{ij} - \mu_{kj})^2$$

- Also, the new centres for the groups are the mean of the values assigned to that group.
- The choice of dissimilarity and the choice of centre are related.

Extensions of k-Means Clustering

- Alternatives to k-means clustering could use a different measure of dissimilarity.
- For example, we could assign observation $\mathbf{x}_i$ to group k so that $d(\mathbf{x}_i, \mu_k)$ is minimized, but where $d(\mathbf{x}_i, \mu_k)$ relates to an alternative measure of dissimilarity to squared Euclidean distance.
- This can help prevent the algorithm from forming circular clusters.
- The new centres for the groups could then be computed by minimizing,

$$\sum_{i=1}^{N} d(\mathbf{x}_i, \mu_k) I_{ik}$$

  Here $I_{ik} = 1$ if observation i is assigned to cluster k and is 0 otherwise.
- The Partitioning Around Medoids algorithm does this (Kaufman and Rousseau 1990).

Cluster Medoids

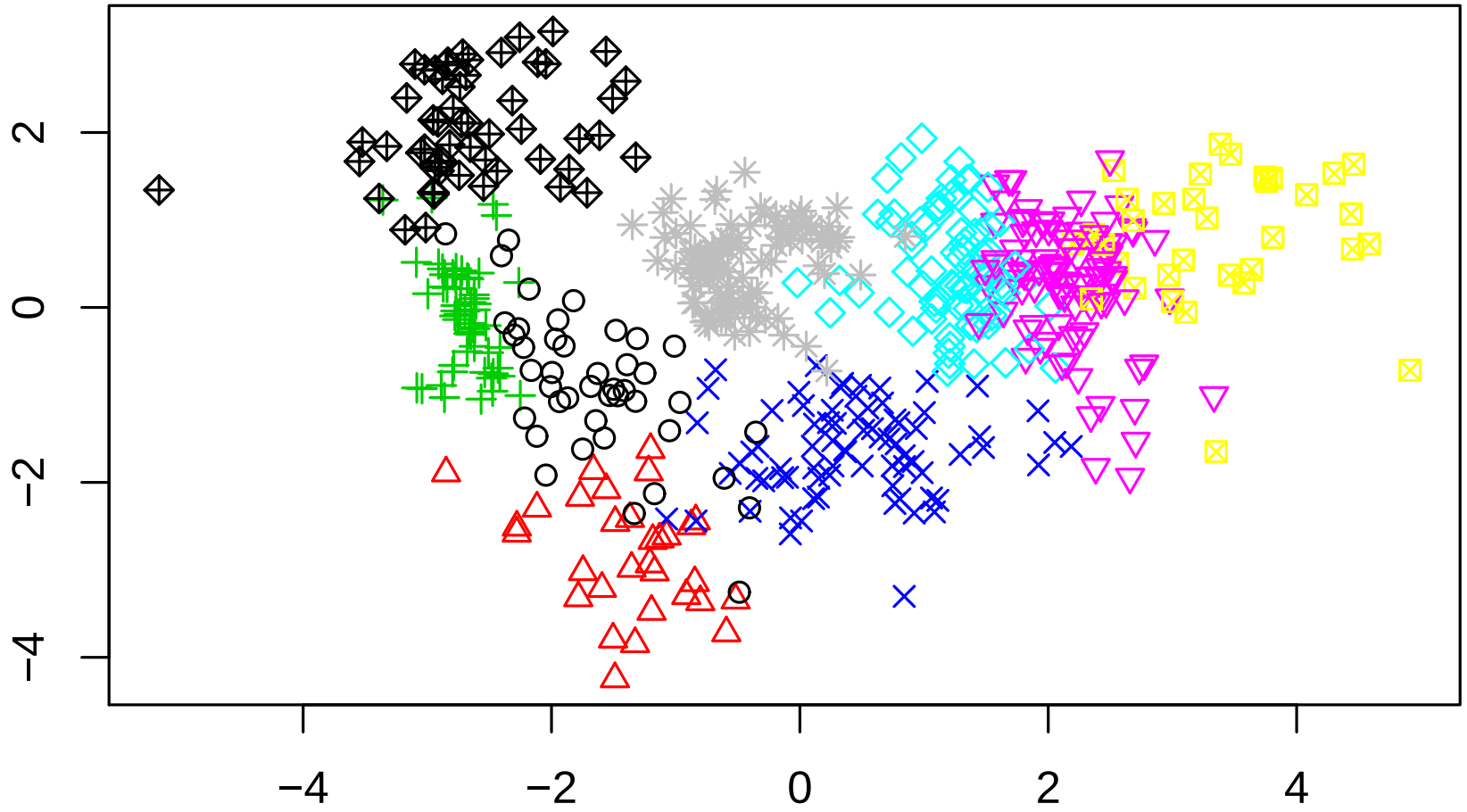
- A medoid is a representative object of a cluster so that its average dissimilarity to all the data points in the cluster is minimal.
- Unlike the means or centroids used in k-means clustering, a medoid has to be an actual data point within the cluster.
- Think along the lines of mean and median in averaging: the median has to be one of the actual numbers, whilst the mean can be something that can't possibly occur (the mean roll of a die is 3.5).
- The use of medoids becomes very useful in applications where a mean or centroid are conceptually difficult to understand, or if they can not even be defined, *e.g.*, 3-D trajectories.

PAM Pseudo-Code

1. Select a dissimilarity metric to be used.
2. Initialize by selecting k of the n data points to be the medoids.
3. Cluster each data point to belong to the same group as the medoid it is closest to under the dissimilarity metric selected.
4. For each medoid $\mathbf{x}^*$:
   - For each non-medoid point $\mathbf{x}$, swap $\mathbf{x}^*$ and $\mathbf{x}$ and compute the total dissimilarity cost of the configuration.
5. Select the configuration with lowest total dissimilarity cost.
6. Repeat 3 to 5 until covergence, *i.e.*, no change in medoids.

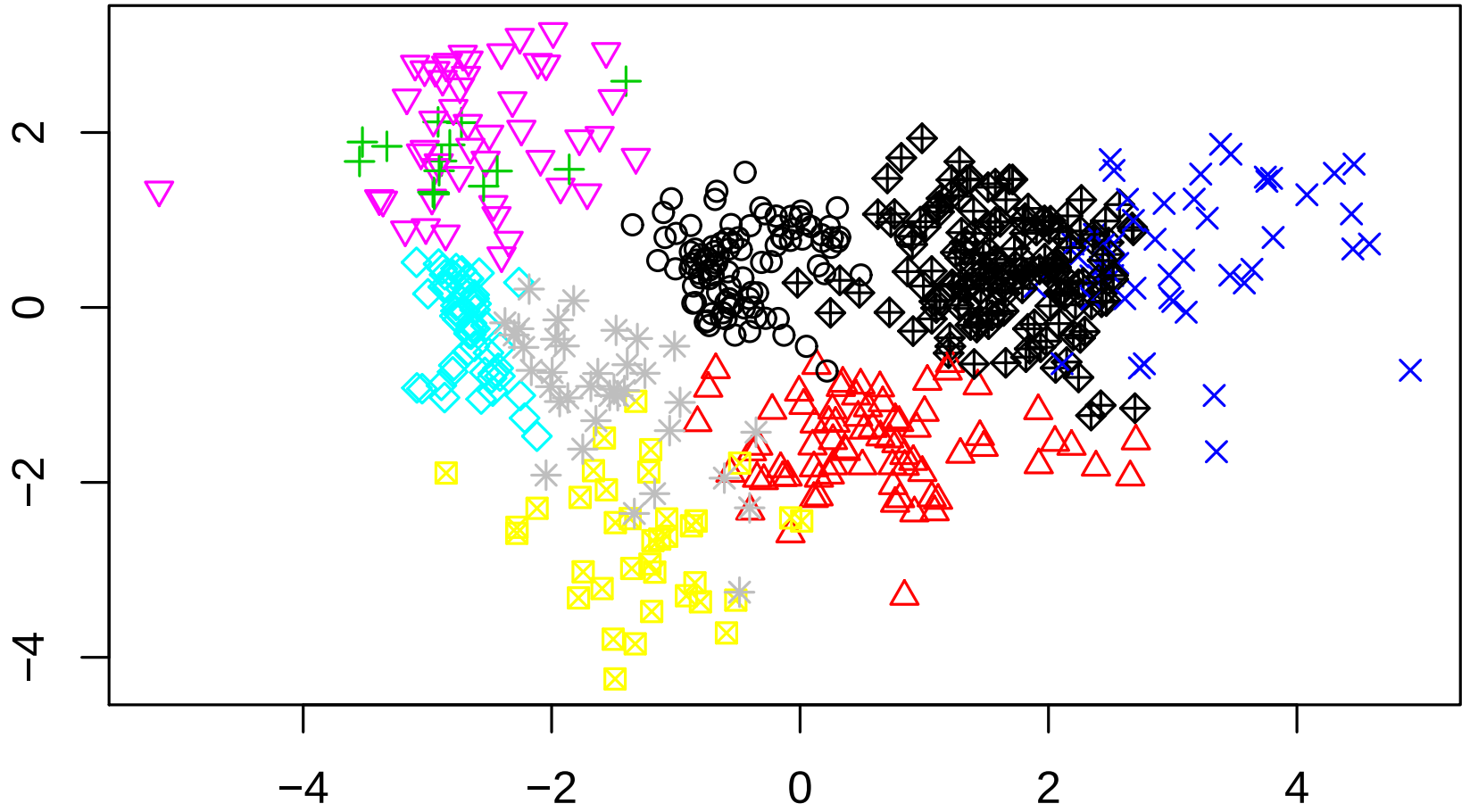
PAM Results

- Clustering the olive oil data using Manhattan dissimilarity gives:



k-means Results

- Clustering the olive oil data using k-means gives:



Agreement?

- A cross tabulation of the cluster memberships shows good agreement.

	8	7	5	2	3	9	4	1	6
1	36	2	2	0	0	0	0	0	3
2	0	31	0	0	0	0	0	0	0
3	0	0	53	0	0	0	0	0	4
4	0	5	0	71	0	0	0	0	0
5	0	0	0	2	0	76	0	0	0
6	0	0	0	3	0	79	17	0	0
7	0	0	0	0	0	0	34	0	0
8	0	0	0	0	0	0	0	98	0
9	0	0	0	0	15	0	0	0	41

- Note, that the columns have been re-ordered to make the matrix as diagonal as possible.

Mixture Models

- Suppose we have data $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{im})$, which is known to arise from one of k populations.
- Within each population (cluster) the data follows a density $f(\mathbf{x}_i|\theta_j)$ for $j = 1, \ldots, k$, where θ_j are the parameters governing population j.
- Suppose the probability that a data point is from population j is π_j.
- Then the data can be modeled using a **mixture model**:

$$P(\mathbf{x}_i) = \sum_{j=1}^{k} P(\mathbf{x}_i \in j) P(\mathbf{x}_i | \mathbf{x}_i \in j)$$

$$P(\mathbf{x}_i) = \sum_{j=1}^{k} \pi_j f(\mathbf{x}_i|\theta_j)$$

- Mixture models can be used to form a model based clustering technique.

Mixture models: $m = 1$

- The π_j values are called the *mixing proportions*, and $f(\cdot|\theta_j)$ is known as the j-th *component density*.
- These models offer good modeling flexibility by allowing both k and the model parameters within each population to vary.
- One common form of mixture model is the mixture of normals, where each component density is a normal density (or multivariate normal density). In the univariate case this is:

$$P(x_i) = \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(x_i - \mu_j)^2}{2\sigma_j^2} \right\}$$

- This assumes that the data within each population is normal with mean μ_j and variance σ_j^2.
- This is a similar idea as LDA and QDA, and we could constrain groups to have the same variance or allow the variance to vary across groups.

Cluster shapes

- Remember that k-means clustering looks for circular clusters.
- LDA fits ellipses of the same shape and direction (equal covariance matrix assumption).
- QDA allows for different covariance matrices between groups, *i.e.*, different shapes and directions.
- Model-based clustering allows different shapes and directions, as well as a varying number of clusters k.

Multivariate Mixture models: $m > 1$

- The normal mixture model can be easily extended to a multivariate mixture model.
- In this case it is assumed that the data within group j follows a multivariate normal distribution with mean μ_j and covariance matrix Σ_j.
- Again we could constrain the covariance matrix in different ways to allow modeling flexibility.

Decomposing covariance matrices

- We have already seen (in PCA) how we can decompose the covariance matrix using an eigenvalue/eigenvector decomposition. We could decompose the covariance matrix Σ as:

$$\Sigma = \lambda \mathbf{D}\mathbf{A}\mathbf{D}^T$$

Here

$$\begin{aligned} \lambda &= \text{a constant} \\ \mathbf{D} &= \text{orthogonal matrix of eigenvectors} \\ \mathbf{A} &= \text{diagonal matrix with entries proportional to eigenvalues.} \end{aligned}$$

- Proof omitted, but it is this result which drives the other proofs that were omitted when covering PCA.

Decomposing Covariance Matrices

- We have also seen that the contours of the density for a multivariate normal form ellipses, the shape of which being controlled by the covariance matrix Σ.
- The various elements of the decomposition of Σ control different aspects of the shape:
  - $\mathbf{A}$ controls the shape of the ellipse
  - $\mathbf{D}$ controls the orientation of the ellipse
  - λ controls the size of the ellipse.
- If $\mathbf{A} = \mathbf{D} = \mathbf{I}$ then the ellipse becomes a circle.
- If $\mathbf{D} = \mathbf{I}$ and $\mathbf{A}$ is unconstrained, then the ellipse becomes aligned with the axes.
- When we move to mixtures of normals the flexibility increases further. This is the basis for model-based clustering.

Mechanism of the Method

- The main aim is to find clusters in the data when k is unknown (unsupervised).
- The general idea is to fit a mixture of normals to the data set for a range of possible values for k and for a range of different possibilities for the Σ's.
- Model fitting uses a maximum likelihood approach via an algorithm known as the Expectation-Maximization (EM) algorithm (this is generally what happens in the k-means and PAM algorithms).
- For a given cluster number k and a likelihood model $L(\theta|\mathbf{X}, \mathbf{z})$ for the probability of parameters θ given data set $\mathbf{X}$ and cluster assignment $\mathbf{z}$, the pseudo-EM algorithm iterates between the following:
  - E-step: Find the expected value of θ as a function of $\mathbf{z}$ using the given likelihood model.
  - M-step: Change cluster assignment $\mathbf{z}$ to maximise the expected likelihood from the E-step.

How to choose the optimum model?

- One way to choose the optimal model is to use the **Bayesian Information Criterion**:

  $$\text{BIC} = -(2 \times \text{maximized likelihood of the data}) + (\log N \times \#\text{ of parameters})$$

- This criterion is calculated for each different type of model fitted, and the model with the smallest value is deemed optimal.
- Note that the first term in the criterion rewards good model fit, whilst the second term penalizes models for having a large number of parameters.
- We will not go into the details of how to perform this technique, rather it is sufficient you know of its existence.

Going Further...

- Alternatively a Bayesian approach can be employed though approximation by simulation, but then we really are moving well beyond what can be expected of this course.
- Model-based clustering will return the optimal number of groups in our data and also indicate the optimal constrained decomposition of the Σ's (we can even incorporate k into the likelihood model).
- It can also return estimates of the group membership of each data point and an estimate of the uncertainty in the group assignment.
- Of course all of this is a live field of active research.