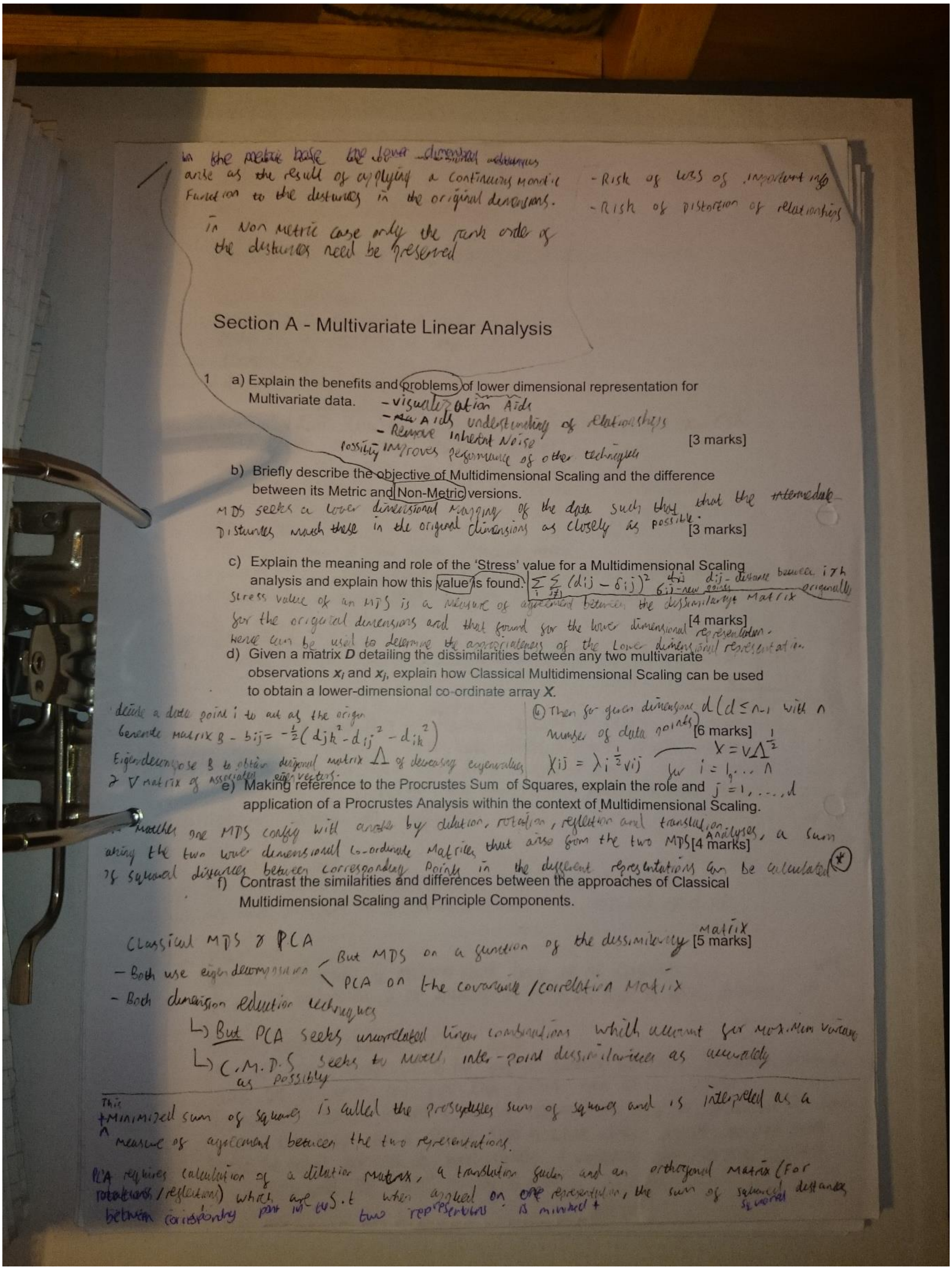


In the metric case the lower dimensional distances arise as the result of applying a continuous monotic function to the distances in the original dimensions.

In Non Metric case only the rank order of the distances need be preserved

- Risk of loss of important info
- Risk of distortion of relationships

Section A - Multivariate Linear Analysis

1 a) Explain the benefits and problems of lower dimensional representation for Multivariate data. [3 marks]

- visualization Aids
- ~~[illegible]~~ Aids understanding of relationships
- Remove inherent Noise

Possibly improves performance of other techniques

b) Briefly describe the objective of Multidimensional Scaling and the difference between its Metric and Non-Metric versions. [3 marks]

MDS seeks a lower dimensional mapping of the data such that ~~they~~ that the intermediate Distances match these in the original dimensions as closely as possible.

c) Explain the meaning and role of the 'Stress' value for a Multidimensional Scaling analysis and explain how this value is found. [4 marks]

$\boxed{\sum_i \sum_{j \neq 1} (d_{ij} - \delta_{ij})^2}$ δ_{ij} - new points d_{ij} - distance between i & h originally

Stress value of an MDS is a measure of agreement between the dissimilarity Matrix for the original dimensions and that found for the lower dimensional representation. Hence can be used to determine the appropriateness of the Lower dimensional representation.

d) Given a matrix ***D*** detailing the dissimilarities between any two multivariate observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, explain how Classical Multidimensional Scaling can be used to obtain a lower-dimensional co-ordinate array ***X***. [6 marks]

decide a data point i to act as the origin

Generate matrix B - $b_{ij} = -\frac{1}{2}(d_{jk}^2 - d_{ij}^2 - d_{ik}^2)$

Eigendecompose B to obtain diagonal matrix Λ of decreasing eigenvalues & V matrix of associated eigenvectors.

(6) Then for given dimension d ($d \leq n-1$ with n number of data points)

$X_{ij} = \lambda_i^{\frac{1}{2}} v_{ij}$ for $i = 1, \dots n$, $j = 1, \dots, d$ — $X = V\Lambda^{\frac{1}{2}}$

e) Making reference to the Procrustes Sum of Squares, explain the role and application of a Procrustes Analysis within the context of Multidimensional Scaling. [4 marks]

[illegible] matches one MDS config with another by dilution, rotation, reflection and translation. Analyses, a sum arising the two lower dimensional co-ordinate Matrices that arise from the two MDS of squared distances between corresponding Points in the different representations can be calculated (*)

f) Contrast the similarities and differences between the approaches of Classical Multidimensional Scaling and Principle Components. [5 marks]

Classical MDS & PCA

- Both use eigendecomposition
  - But MDS on a function of the dissimilarity Matrix
  - PCA on the covariance / correlation Matrix
- Both dimension reduction techniques
  - ↳ But PCA seeks uncorrelated linear combinations which account for maximum variance
  - ↳ C.M.D.S seeks to model inter-point dissimilarities as accurately as possibly

(*) This minimized sum of squares is called the procrustes sum of squares and is interpreted as a measure of agreement between the two representations.

PCA requires calculation of a dilation matrix, a translation vector and an orthogonal Matrix (For rotations / reflections) which are s.t when applied on one representation, the sum of squared distances between corresponding point in the two representations is minimised +

MLA

7/05/15.

HIERARCHICAL CLUSTERING

Establish if there is a group structure in the data set. → How many groups? structure?

Similarity / Dissimilarity

- Want to place observations in groups according to their similarity
- $d(x,y) \geq 0$ and $d(x,y) = 0$ if $x = y$
- $d(x,y) = d(y,x)$
- $d(x,z) \leq d(x,y) + d(y,z)$

Euclidean: $\sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$ $\qquad x_i^T = (x_{i1}, x_{i2}, \dots x_{im})$ $x_{ik} \in \mathbb{R}$

Absolute Distance (Manhattan): $\sum_{k=1}^{m} |x_{ik} - x_{jk}|$

Maximum: $\max_{k \in \{1,2 \dots m\}} |x_{ik} - x_{jk}|$

Minkowski: $[\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p]^{1/p}$ $(p \geq 1)$

Standardisation

- Need to be aware of scaling before dissimilarity matrix
- Variables need to be scaled otherwise variable with larger variance will figure most prominently.
- standardise by dividing by variance, all have variance of 1 then

Binary Data

- Consider dissimilarity by looking at cross tabulation of 0's and 1's

		Point j 1	0	
	1	a	b	a+b
point i	0	c	d	c+d
		a+c	d+b	m=a+b+c+d

- Simple matching (Hamming) $1 - \frac{a+d}{a+b+c+d}$ proportion of variables in agreement.
- Jaccard: $1 - a/a+b+c$ ignore double absence, as may be redundant variable
- Kulczynski: $1 - \frac{1}{2}(a/a+b + a/a+c)$ Average of ratios of agreement from two samples
- Czekanowski: $1 - 2a/2a+b+c$ More emphasis on double presence than double absence.

Choice depends on application

Categorical Data

- Use simple matching - count number of terms that are different.
- Eg (Male, Brown, Brown, Scurvy) (Female, Brown, Green, Third) differ in 3 → d = 3

Mixed data

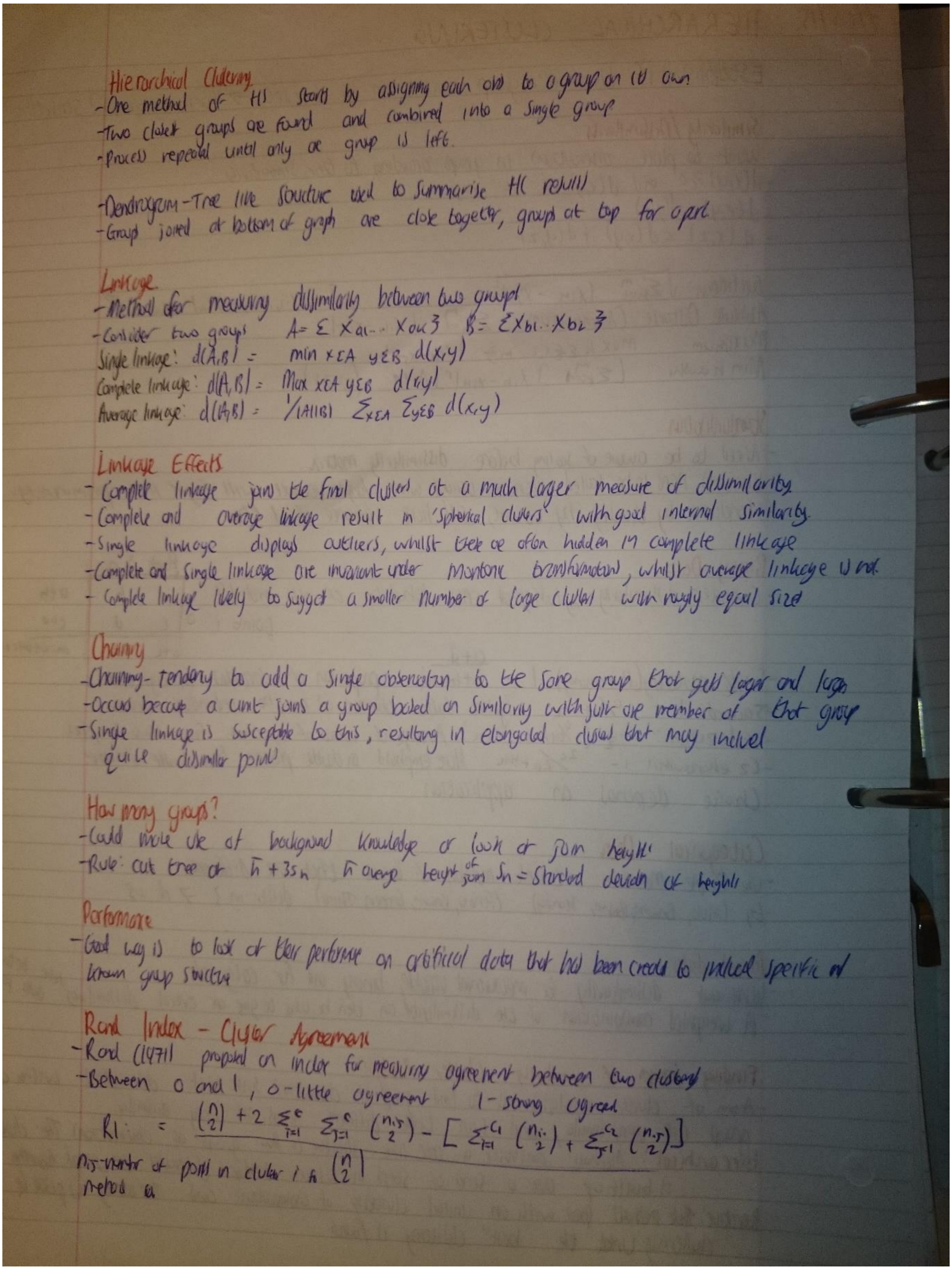
- Work out dissimilarity for measurement variable, binary and for categorical
- A weighted combination of the dissimilarities can then be used to give an overall dissimilarity value between data points

Finding Groups of Similarity - Cluster Analysis

- Aim of cluster analysis is to find groups of observations such that observations within a group are very similar, and such that different groups are very dissimilar.

Hierarchical: Methods construct a tree like structure to show groups of observations. The clustering is built up over a series of steps in which similar observations are joined together

Iterative: These methods start with an initial clustering of observations and iteratively update the clustering until the "best" clustering is found



Hierarchical Clustering

- One method of HC starts by assigning each obs to a group on its own
- Two closest groups are found and combined into a single group
- Process repeated until only one group is left.

- Dendrogram - Tree like structure used to summarise HC results
- Groups joined at bottom of graph are close together, groups at top far apart

Linkage

- Method for measuring dissimilarity between two groups
- Consider two groups $A = \{x_{a1} \dots x_{a2}\}$ $B = \{x_{b1} \dots x_{b2}\}$

Single linkage: $d(A,B) = \min_{x \in A\, y \in B} d(x,y)$

Complete linkage: $d(A,B) = \max_{x \in A\, y \in B} d(x,y)$

Average linkage: $d(A,B) = 1/|A||B| \sum_{x \in A} \sum_{y \in B} d(x,y)$

Linkage Effects

- Complete linkage joins the final clusters at a much larger measure of dissimilarity
- Complete and average linkage result in 'Spherical clusters' with good internal similarity.
- Single linkage displays outliers, whilst these are often hidden in complete linkage
- Complete and Single linkage are invariant under monotonic transformations, whilst average linkage is not.
- Complete linkage likely to suggest a smaller number of large clusters with roughly equal size

Chaining

- Chaining - tendency to add a single observation to the same group that gets larger and larger
- Occurs because a unit joins a group based on similarity with just one member of that group
- Single linkage is susceptible to this, resulting in elongated clusters that may include quite dissimilar points

How many groups?

- Could make use of background knowledge or look at join heights
- Rule: cut tree at $\bar{h} + 3s_h$ $\bar{h}$ average height of join s_h = Standard deviation of heights

Performance

- Good way is to look at their performance on artificial data that has been created to include specific and known group structure

Rand Index - Cluster Agreement

- Rand (1971) proposed an index for measuring agreement between two clusterings
- Between 0 and 1, 0 - little agreement 1 - strong agreement

$$RI = \frac{\binom{n}{2} + 2\sum_{i=1}^{c}\sum_{j=1}^{c}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}\right]}{\binom{n}{2}}$$

n_{ij} - number of points in cluster i for method a

MLA 3

HIEARCHICAL CLUSTERING.

- n_{ij} - number of points that are in cluster i for method A and cluster j for method B
- c_1 is number of clusters for method A
- c_2 number of clusters for method B
- $n_{\cdot j} = \sum_{i=1}^{c_1} n_{ij}$. $n_{i\cdot} = \sum_{j=1}^{c_2} n_{ij}$. $n = n_{\cdot\cdot} = \sum_{i=1}^{c_1} \sum_{j=1}^{c_2} n_{ij}$

Method B

		cluster 1	cluster 2	cluster c_2	
Method A.	cluster 1	n_{11}	n_{12}	n_{1c_2}	$n_{1\cdot}$
	cluster 2	n_{21}	n_{22}	n_{2c_2}	$n_{2\cdot}$
	cluster c_1	$n_{c_1 1}$	$n_{c_1 2}$	$n_{c_1 c_2}$	$n_{c_1\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot c_2}$	$n_{\cdot\cdot}$

- RI tends to give quite large values even when clustering methods are in substantial disagreement
- Considering a distribution for assigning points to clusters under the condition that cluster sizes remain unchanged

$$\text{Adjusted Rand} = \frac{\binom{n}{2} \sum_{i=1}^{c_1} \sum_{j=1}^{c_2} \binom{n_{ij}}{2} - \sum_{i=1}^{c_1} \binom{n_{i\cdot}}{2} \sum \binom{n_{\cdot j}}{2}}{\frac{1}{2}\binom{n}{2} \left[\sum_{i=1}^{c_1} \binom{n_{i\cdot}}{2} + \sum_{j=1}^{c_2} \binom{n_{\cdot j}}{2}\right] - \sum_{i=1}^{c_1} \binom{n_{i\cdot}}{2} \sum_{j=1}^{c_2} \binom{n_{\cdot j}}{2}}$$

$$\text{Arlies Fm:} \quad \frac{\text{Rand Index} - \text{Expected Rand Index}}{\text{Max Rand Index} - \text{Expected Rand Index}}$$

MLA

08/05/15

CLUSTERING (KNN)

K-Nearest Neighbours (KNN)

- A non parametric/distribution free method of assigning group membership
- i.e. makes no assumption about spread of data within each class
- Consequence is that class assignment is fixed, with no measure of uncertainty concerning any particular assignment
- Classification techniques that do make distributional assumptions allow quantification of the uncertainty in group measurement.

- KNN looks at k closest points of known origin to the point of unknown origin
- Point is then classified as belonging to the group which contains the most of these k-points
- Results ARE NOT INVARIANT TO SCALING of the original variables (standardise) nor to the method in which distance is calculated, as these are likely to change the nearest neighbours for any particular point.

- KNN method classifies new observation as belonging to the class that was most prevalent in those k labelled neighbours.

Choosing K

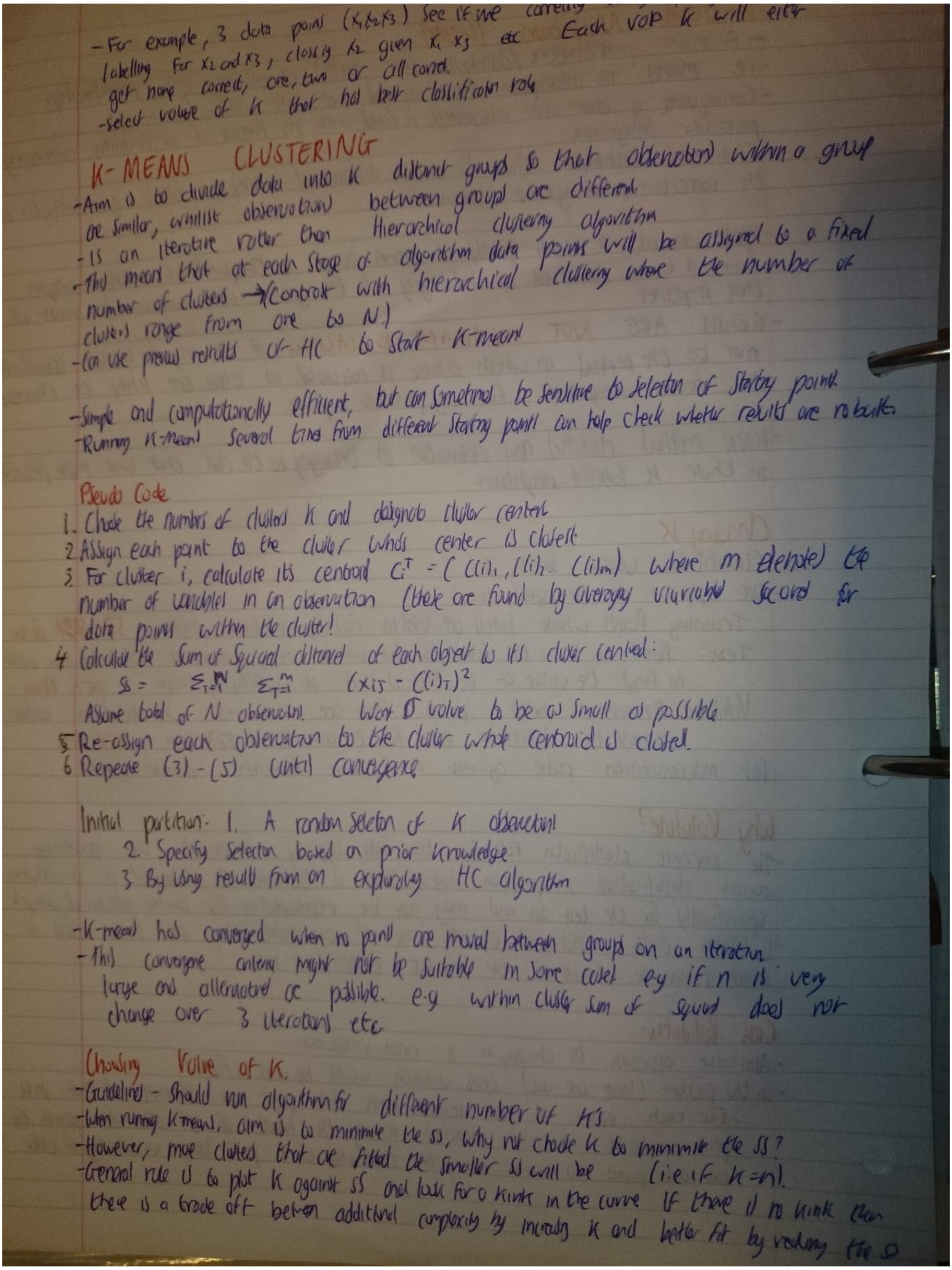
- Classification varies with k
- One approach is to split data:
  - Training: Points whose labels are used to classify unlabelled points 50% Data
  - Test: Points we know the labels for but which are considered unlabelled in order to find the value of k that is best at classifying them 25% Data
  - Validation: Remaining labelled points that are considered unlabelled in order to estimate the classification error for the best k in Test step 25% Data
- Plot misclassification rate against k - choose k with lowest y value

Why Validate?

- The correct classification rate for test data typically overestimates the percentage correct classifications in validation data. This is because the value of k is chosen specifically for the test set and may not be representative for another unlabelled sample
- Validation data is not used at any stage of the model fitting and hence offers a more reasonable estimate of the correct classification rate

Cross Validation

- Alternative approach to choosing k is cross validation
- In this problem (leave-one-out) cross validation would be used as follows:
  - For each value of k remove each data point and determine if that data point would be correctly classified knowing the labels of all other data points. Eg. If 100 data points this means making 100 classifications of 1 point based on labelling of other 99



- For example, 3 data points (x_1, x_2, x_3) See if we correctly Each value k will either labelling for x_2 and x_3, classify x_2 given x_1 x_3 etc get none correct, one, two or all correct.
- select value of k that has best classification rate

K-MEANS CLUSTERING

- Aim is to divide data into K distinct groups so that observations within a group are similar, whilst observations between groups are different
- Is an iterative rather than Hierarchical clustering algorithm
- This means that at each stage of algorithm data points will be assigned to a fixed number of clusters → (Contrast with hierarchical clustering where the number of clusters range from one to N)
- Can use previous results of HC to start K-means

- Simple and computationally efficient, but can sometimes be sensitive to selection of starting points
- Running K-means several times from different starting points can help check whether results are robust.

Pseudo Code

```
1. Choose the numbers of clusters k and designate cluster centers
2 Assign each point to the cluster whose center is closest
3. For cluster i, calculate its centroid $C_i^T = (C(i)_1, C(i)_2 \; C(i)_m)$ where m denotes the
   number of variables in an observation (these are found by averaging variable scores for
   data points within the cluster)
4 Calculate the sum of squared distances of each object to its cluster centroid:
      $SS = \Sigma_{T=1}^{N} \; \Sigma_{T=1}^{m} \; (x_{iJ} - C(i)_T)^2$
   Assume total of N observations. Want SS value to be as small as possible
5 Re-assign each observation to the cluster whose centroid is closest.
6 Repeat (3) - (5) until convergence
```

Initial partition. 1. A random selection of k observations
2. Specify selection based on prior knowledge
3. By using results from an exploratory HC algorithm

- K-means has converged when no points are moved between groups on an iteration
- This convergence criteria might not be suitable in some cases e.g. if n is very large and alternatives are possible. e.g. within cluster sum of squares does not change over 3 iterations etc

Choosing Value of K.

- Guidelines - Should run algorithm for different number of K's.
- When running K-means, aim is to minimise the SS, why not choose k to minimise the SS?
- However, more clusters that are fitted the smaller SS will be (i.e if k=n).
- General rule is to plot K against SS and look for a kink in the curve If there is no kink then there is a trade off between additional complexity by increasing k and better fit by reducing the SS

5/15

CLUSTERING

- K-means clustering looks for circular clusters
- Can partitional plane by drawing line which are equidistance from the means to create regions within plane

- K means can give different answers when initialised at different starting values
- Algorithm does not always find the minimum value for TWSS
- Choose total within sum of squares which is smallest

Model Based - Clustering

- Makes use of - Statistical models → parametric in nature

- K means assigns obs to the group which has closest centre in terms of squared euclidean distance.
- A obs $X_i^T (X_{i1} \ldots X_{im})$ is assigned to group k so that $d(x, \mu_k)$ is minimised where: $d(x_i, \mu_k) = \sum_{j=1}^{m} (x_{ij} - \mu_{kj})^2$
- Also, new ~~[illegible]~~ centres for groups are mean of the values assigned to their group
- Choice of dissimilarity and center are related

- Can use different dissimilarity other than Euclidean, can help prevent forming circular clusters
- New centers could be computed by minimising $\sum_{i=1}^{N} d(x_i, \mu_k) I_{ik}$
  Here $I_{ik} = 1$ if obs i is assigned to cluster k and 0 otherwise
- Partitioning around medoids PAM does this

Cluster Medoid

- A medoid is a representative object of a cluster so that its average dissimilarity to all the data points in cluster is minimal
- Unlike mean or centroid, used in k-means a medoid has to be an actual data point

Partitioning Around Medoid (PAM) Pseudo Code

1. Select a dissimilarity metric to be used
2. Initialise 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 and the dissimilarity metric selected.
4. For each medoid x^*: - For each non medoid point x, swap x^* and x and compute total dissimilarity cost of the configuration
5. Select the configuration with lowest total dissimilarity cost
6. Repeat 3-5 until convergence i.e no change in medoids

Mixture Models

- Suppose we have data $X_i^T = (X_{i1}, X_{i2} \ldots X_{im})$ which is known to arise from one of K populations
- Within each population (cluster) the data follow a density $f(x_i | \theta_j)$ for $j = 1 \ldots K$ where θ_j are the params 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:

$$p(x_i) = \sum_{j=1}^{k} p(x_i \in j) P(x_i | x_i \in j)$$

$$p(x_i) = \sum_{j=1}^{k} \pi_j f(x_i | \theta_j)$$

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

Mixture Models m=1

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

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

- Assumes data within each population is normal with mean μ_j and variance σ_j^2.
- Similar idea to LDA QDA and we could constrain groups to have the same variance or allow the variance to vary across groups

Cluster shapes

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

- Normal mixture model can be easily extended to a multivariate mixture model
- Assumes that data within group j follows a multivariate normal dist with mean μ_j and covariance matrix Σ_j.
- Could constrain the cov matrix in different ways to allow modeling flexibility

- Can decompose covariance matrix as: $\Sigma = \lambda D A D^T$

λ = a constant D = orthogonal matrix of eigen vectors

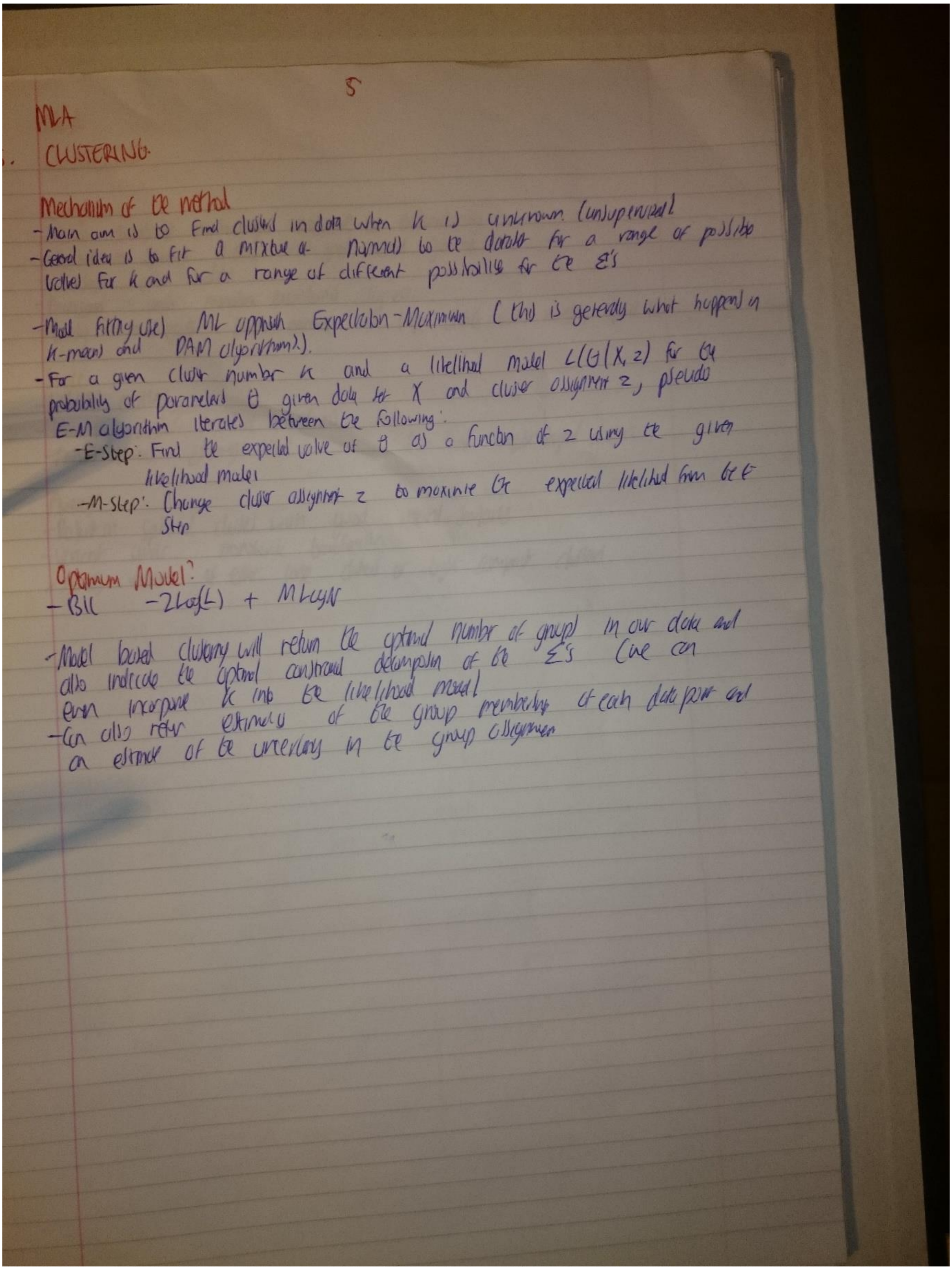
A = diagonal matrix with entries proportional to eigenvalues

A - controls shape of the ellipse

D - controls the orientation of the ellipse

λ - controls the size of the ellipse

- If $A = D = I$ then ellipse becomes circle
- If $D = I$ and A is unconstrained, ellipse becomes aligned with the axes
- When we move to mixtures of normals, flexibility increases



MLA

CLUSTERING.

Mechanism of the method

- Main aim is to find clusters in data when k is unknown (unsupervised)
- General idea is to fit a mixture of normals to the data for a range of possible values for k and for a range of different possibilities for the Σ's

- Model fitting uses ML approach Expectation-Maximum (this is generally what happens in k-means and PAM algorithms).
- For a given cluster number k and a likelihood model $L(\theta|X,z)$ for the probability of parameters θ given data set X and cluster assignment z, pseudo E-M algorithm iterates between the following:
  - E-Step: Find the expected value of θ as a function of z using the given likelihood model
  - M-Step: Change cluster assignment z to maximise the expected likelihood from the E step

Optimum Model?

- BIC $-2\log(L) + M\log N$

- 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)
- Can also return estimates of the group membership of each data point and an estimate of the uncertainty in the group assignment

MLA

10/05/15. CLUSTERING

Complete

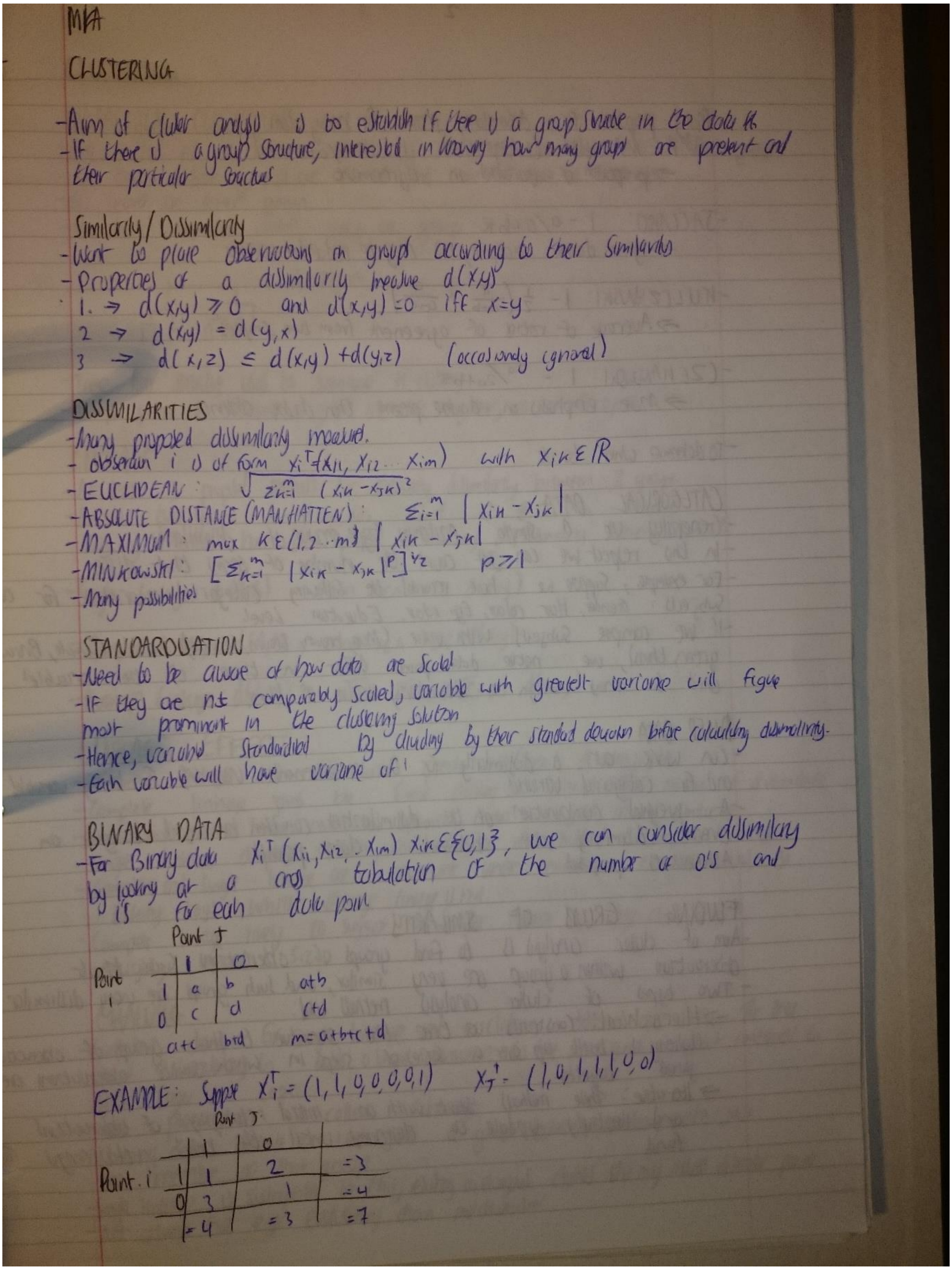
- Joins clusters at much larger measure of dissimilarity
- Result in spherical clusters with good internal synergy
- Outliers often hidden
- Invariant under monotonic transforms log etc.
- likely to suggest smaller number of large clusters with roughly equal size

Single.

- Identifies outliers.
- Invariant under monotonic transforms
- Tends toward chaining effect. Can find irregular-shaped clusters

Average

- Result in spherical clusters with good internal similarity
- Variant under monotonic transformation
- Avoids extremes of either large clusters or tight compact clusters



MVA

CLUSTERING

- Aim of cluster analysis is to establish if there is a group structure in the data th.
- If there is a group structure, interested in knowing how many groups are present and their particular structures

Similarity / Dissimilarity

- Want to place observations in groups according to their similarity
- Properties of a dissimilarity measure $d(x,y)$
1. → $d(x,y) \geqslant 0$ and $d(x,y) = 0$ iff $x = y$
2. → $d(x,y) = d(y,x)$
3. → $d(x,z) \leq d(x,y) + d(y,z)$ (occasionally ignored)

DISSIMILARITIES

- Many proposed dissimilarity measures.
- observation i is of form $x_i^T = (x_{i1}, x_{i2} \dots x_{im})$ with $x_{ik} \in \mathbb{R}$
- EUCLIDEAN: $\sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$
- ABSOLUTE DISTANCE (MANHATTEN): $\sum_{i=1}^{m} | x_{ik} - x_{jk} |$
- MAXIMUM: $\max_{k \in \{1,2 \dots m\}} | x_{ik} - x_{jk} |$
- MINKOWSKI: $[\sum_{k=1}^{m} | x_{ik} - x_{jk} |^p]^{1/2}$ $p \geqslant 1$
- Many possibilities

STANDARDISATION

- Need to be aware of how data are scaled
- If they are not comparably scaled, variable with greatest variance will figure most prominent in the clustering solution
- Hence, variables standardised by dividing by their standard deviation before calculating dissimilarity.
- Each variable will have variance of 1

BINARY DATA

- For Binary data $x_i^T (x_{i1}, x_{i2} \dots x_{im})$ $x_{ik} \in \{0,1\}$, we can consider dissimilarity by looking at a cross tabulation of the number of 0's and 1's for each data point

		Point j		
		1	0	
Point i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	m=a+b+c+d

EXAMPLE: Suppose $x_i^T = (1,1,0,0,0,0,1)$ $x_j^T = (1,0,1,1,1,0,0)$

		Point j		
		1	0	
Point. i	1	1	2	=3
	0	3	1	=4
		=4	=3	=7

Many proposals for dissimilarity measures for binary data.

- SIMPLE MATCHING (HAMMING): $1 - a+b/a+b+c+d$
  - ⇒ proportion of variables in agreement
- JACCARD: $1 - a/a+b+c$
  - ⇒ Ignore double absence, a) may be redundant variable
- KULCZYNSKI: $1 - \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$
  - ⇒ Average of ratios of agreement from two samples
- CZEKANOWSKI: $1 - {}^{2a}/_{2a+b+c}$
  - ⇒ More emphasis on double presence than double absence
- To determine which to use

CATEGORICAL DATA

- Generally use a simple matching type measure of dissimilarity
- In this respect we can just count the number of terms that differ.
- For example, Suppose we have recorded the following categorical variables for two Subjects: Gender, Hair colour, Eye colour, Education Level
- If we compare Subjects with values (Male, brown, Brown, Secondary) and (Female, Brown, green, third), we notice data points differ in three of the variables and so may assign a dissimilarity value of three

MIXED DATA

- Can work out a dissimilarity for the measurement variables, for binary variables and for categorical variables
- A weighted combination of the dissimilarities can then be used to give an overall dissimilarity value between data points
- Alternatives proposed

FINDING GROUPS OF SIMILARITY

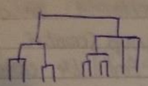
- Aim of cluster analysis is to find groups of observations such that observations within a group are very similar, and such groups are very dissimilar
- Two types of cluster analysis methods used:
  - ⇒ Hierarchical: Constructs a tree like structure to show groups of observations clustering is built up over a series of steps in which similar observations are joined together.
  - ⇒ Iterative: These methods start with an initial clustering of observations and iteratively update the clustering until the "best" clusterings found

/04/15.

CLUSTERING

HIEARCHICAL CLUSTERING

- One method of H clustering starts by assigning each observation to its own group
- The two closest groups are found and combined into a single group
- This leaves one fewer group
- Process is repeated until only one group is left.



Dendogram

- Tree like structure used to summarise H-clustering
- Groups joined at bottom of graph are close together, at top - far apart

LINKAGE

- At least 3 proposed methods for measuring dissimilarity between 2 groups.
- Consider two groups $A = \{x_{a1}, x_{a2} \dots x_{ak}\}$ $B = \{x_{b1}, x_{b2} \dots x_{bl}\}$
- Following Methods have been proposed for measuring the dissimilarity between A and B.

- Single Linkage $d(A,B) = \min_{x \in A \; x \in B} d(x,y)$

- Complete Linkage: $d(A,B) = \max_{x \in A \; x \in B} d(x,y)$.

- Average Linkage: $d(A,B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{x \in B} d(x,y)$

LINKAGE EFFECTS

- Different linkages and dissimilarities create different dendrographs.
- Complete linkage joins the final clusters at a much higher measure of dissimilarity
- Complete and average linkage result in 'spherical clusters' with good internal similarity
- Single linkage displays outliers, whilst these are often hidden in complete linkage
- Complete and single linkage are invariant under monotonic transformations, eg taking logs, whilst average linkage is not.
- Complete linkage likely to suggest a smaller number of large clusters with roughly equal size

CHAINING

- Consider dendro for single linkage and euclidean ~~distance~~. dissimilarity. The tree shows a tendency to add a single observation to the same group that continues to get larger and larger.
- This phenomenon is called "chaining"
- Chaining occurs because a unit joins a group based on similarity with just one member of that group
- Single linkage is susceptible to this, ending in elongated clusters that may include dissimilar points
- Not always bad e.g. evolutionary chain mechanism

HOW MANG GROUPS?

- One possibility is to look at the dendrogram for joins that happen at very large height values
- This is because height on dendrogram is interpretable through the method of linkage used
- Suggested rule is to cut the tree at $\bar{h}+3s_h$, where $\bar{h}$ is the average height of the joins and S_h is the standard deviation of the heights

PERFORMANCE

- Good way to assess performance is to look at their performance on artificial data that has been created to include specific and known group structure.
- Allows a test to determine if the method did indeed find the correct structure when it is known to exist.

CLUSTER AGREEMENT: CROSS TABULATION

- Suppose two different clustering methods are applied to the same data
- A cross tabulation of the cluster memberships from the two methods permits a comparison of results

		Method B		
		C1	C2	
Eg Method A	C1	10	20	4.
	C2	60	15	75
		70	45	11

RAND INDEX

- Rand Index is a number between 0 and 1 with 0 representing little agreement and 1 representing strong agreement
- Formula $$RI = \frac{\binom{n}{2} + 2\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}\right]}{\binom{n}{2}}$$

n_{ij} is number of points that are in cluster i for method A and cluster j for method B,

- C_1 is number of clusters for Method A, C_2 number of clusters in Method B
- $n_{\cdot j} = \sum_{i=1}^{c_1} n_{ij}$ $\quad n_{i\cdot} = \sum_{j=1}^{c_2} n_{ij}$
- $n = n_{\cdot\cdot} = \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} n_{ij}$.

		Method B				
		C1	C2	…	C_{c_2}	
Method A	C1	n_{11}	n_{12}	…	n_{1c_2}	
	C2	n_{21}	n_{22}	…	n_{2c_2}	$n_{1\cdot}$
	⋮	⋮	⋮	⋱	⋮	$n_{2\cdot}$
	C_1	$n_{c_1 1}$	$n_{c_1 2}$	…	$n_{c_1 c_2}$	⋮
		$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c_2}$	$n_{c_1\cdot}$

MLA

/15. CLUSTERING.

RAND INDEX: PROBLEMS.

- Rand Index tends to give quite large values even when clustering method are in Substantial disagreement.
- Even a random assignment of points to clusters can lead to large Rand Index value
- Adjusted rand index - in order to account for agreement by chance. Ad this by considering a distribution for assigning points to clusters under the condition that cluster sizes remained unchanged

Adjusted Rand:

$$\frac{\binom{n}{2}\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2}\sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}}{\left(\frac{1}{2}\right)\binom{n}{2}\left[\sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2}+\sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}\right] - \sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2}\sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}}$$

Arises from: $\text{Adjusted Rand} = \dfrac{\text{Rand Index} - \text{Expected Rand Index}}{\text{Max Rand Index} - \text{Expected Rand Index}}$

Index can be negative but not >1.