MLA

18/04/15. 2013 EXAM PAPER Q2

2 A. HIEARCHICAL CLUSTERING
- Each point starts in it's own cluster.
- Clusters merged by combining closest 2 at each depth
- Closest is defined by dissimilarity matrix - measure for linkage for clusters
- Continue until a single cluster is formed consisting of all data points
- Go back and determine when merging should of stopped. eg rule of avevage join height + 3 * S.d. of the heights or relatively large jump in next merging distance - as seen through dendrograph

ITERATIVE CLUSTERING
- Pre defined number of clusters
- Assign data points to a specific cluster.
- Iterate points between clusters until convergence
- Convergence: repeat of a previous clustering of the data [choose that with minimum internal dissimilarities]
- Examples include K-means, partitioning around medoids.

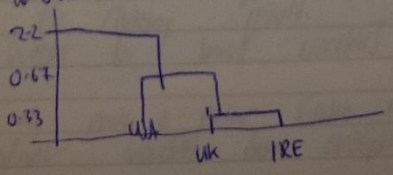
2 B i. Scaling Data
- Important that different variables are scaled, as if they are not, the variable with the greatest variance will figure most prominently in the clustering solution

Subtracting mean
- removes bias
- will effect result ⇒ produce more accurate result

C.

	Ireland	UK	USA	CHINA
Ireland	0	0·33	0·53	1·66
UK	0·33	0	0·67	1·53
USA	0·53	0·67	0	2·20
China	1·66	1·53	2·20	0

D. merge UK and Ireland 0·33.
merge USA and UK at 0·67
China + USA at 2·2.



Cut at 2 groups

H + 3.d of height

MLA

08/05/18 EXAM PAPER 2013 Q2 CLUSTERING

2 A. Hierarchical
- Start with each point of a cluster on it's own
- Clusters are merged by combining closest 2 at each depth
- Closest defined by dissimilarity matrix.
- Continue until a single cluster is formed consisting of all data points
- Go back and determine when merging should of stopped. $\bar{h} + 3s_h$ or when a relatively large jump in next merging distance as seen through dendrogram

Iterative
- Pre defined number of clusters - Specify number of clusters you want before starting
- Assign data point to a specific cluster.
- Iterate and re-assign membership until convergence
- Convergence - repeat of previous clustering of data (choose that with minimal internal dissimilarity)
- Example include: k-means, partitioning around medoids

B i. Effect of Scaling data
- Important that different variables are scaled, if they are not, be variable with greatest variance will figure more prominently in cluster solution
- It can alter distance between points and subsequently affect the algorithm

ii. Subtracting mean vector.
- Has no effect on the clustering. It merely centralises the data.

C. Maximum Dissimilarity

	IRL	UK	USA	CHINA
IRL	0	0·33	0·53	1·66
UK	0·33	0	0·67	1·53
USA	0·53	0·67	0	2·20
CHINA	1·66	1·53	2·20	0

D Dendrograph
Merge UK + IRL at 0·33
Merge USA and UK at 0·67
Merge China + USA at 2·2

E Use $\bar{h} + 3s_h$ or chose point at which relatively large jump occurs in join height.
Suggest two groups

$\bar{h} = \frac{0.33 + 0.67 + 2.2}{3} = \frac{16}{15}$.

Variance = $(0.67 - \frac{16}{15})^2 + (0.33 - \frac{16}{15})^2 + (22 - \frac{16}{15})^2$

$\frac{1024}{5625} + 0.5426 + \frac{289}{225} = 2.409$

$= \frac{\sqrt{2}}{3} - 1.4 = 0.816$

~~$\frac{16}{15} + 3(0.816)$~~ $= 5.26$

$\frac{16}{15} + 3(0.816) =$

$= 3.514$ 2.448

2021A

04/05/15 EXAM PAPER 203 Q1 MDS DAVID WEATHERCOAT

1 A. Benefits and pitfalls of Lower Dimensional Summary
- Lower dimensional summary
- Can visualise Data
- Helps identify relationship in data
- Potentially allows other Multivariate techniques
- Reduced computational requirements
- Removes inherent noise

- Information may be lost
- Data may have to be standardised
- Suitable model may not be available ie too many PCs
- Cannot capture all the data variance only an approximation

B Aim of MDS
MDS seeks to produce a lower dimensional representation of the data such that distance between points i and j in the representation δ_{ij} are close to the dissimilarities between those points d_{ij} for all i, j.

Metric and Non-Metric MDS
Metric - refers to when F is a continuous and monotone function. Preserves the ratio between points
Non-Metric - Only rank order taken into account. Need only obey monotonicity constraint

C STRESS
- Stress is a ~~measure of accuracy~~ degree of correspondence between the distance among points from the MDS map and the original matrix input by user which is used to create original dissimilarities - Measure of inaccuracy or distortion
- Smaller stress = better accuracy

- Stress defined as $\sum_{i=2}^{n} \sum_{i<j} (\delta_{ij} - d_{ij})^2$
- δ_{ij} is distance between i and j in plot and d_{ij} is distance in dissimilarity matrix

- Sammon stress generally preferred as it takes into account the size of distance being approximated
- Stress-1 - used a particular configuration reproduced the observed distance matrix

D MDS Algorithm
1 Obtain the dissimilarities $\{d_{ij}\}$
2 Form B, each element which is given by $b_{ij} = -\frac{1}{2}(d^2_{ij} - d_{i\bullet}^2 - d_{\bullet j}^2)$ with $\bullet$ representing the centroid / origin of all observations
3 Create matrix Λ from the eigenvalues $\lambda_1 \ldots \lambda_{n-1}$ and the matrix V from the associated eigenvectors $V_1 \ldots V_{n-1}$ of B
4 Chose an appropriate number of dimensions q, using a suitable measure

5. The coordinates of the n required points that are used to represent the n observations in d-dimensional space are given by $x_{ij} = \lambda_j^{1/2} v_{ij}$ for $i=1..n$ and $j=1..d$

E. Role of Procrustes Sum of Squares
- Procrustes matches one MDS configuration with another by dilation, rotation, reflection and translation
- Allows us to create several MDS configurations with the same dissimilarity
$R^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} (y_{ij} - x_{ij})^2$
- Lower value indicates a similar measure

F. PCA and MDS
- Same aim - dimension reduction
- Neither make any assumptions about the data
- PCA is not scale invariant and data is generally standardised first.
- MDS does not require standardisation
- MDS uses eigendecomposition of the dissimilarity matrix whereas as PCA uses eigen decomposition of the covariance matrix.
- PCA seeks a linear combination of $x_1..x_n$ which are uncorrelated with high variance
- MDS aims to produce an optimal number of configurations in a smaller number of dimensions - MDS aims to preserve the dissimilarities.

When euclidean distance is used within classical MDS, the resulting lower dimensional co-ordinates are the same as PCA coordinates

MLA

EXAM PAPER 2013 Q1. MDS

1A. BENEFITS/PROBLEMS OF DIMENSION REDUCTION

- Visualisation of data
- Lower dimensional samples
- Helps identify relationships in data
- Potentially assists other clustering techniques

- Information may be lost
- Data may have to be standardised
- Suitable model may not be available, i.e. too many IVAs

B OBJECTIVE OF MDS

- MDS used in pursuit of following form: For a set of dissimilarities between every pair of n items, find a representation of the items in $\mathbb{R}^d$ ($d \leq n$) such that the inter point distances match the original dissimilarities as close as possible
- MDS seeks to produce a lower dimensional representation of the data such that distances between points i and j in the representation, δ_{ij} are close to the dissimilarities between these points d_{ij} for all i, j.

Metric MDS - Different ideas of matching δ_{ij} to $f(d_{ij})$. refers to when f is a continuous and monotone function. e.g. the identity function or a function converting dissimilarities into a distance like form

Non metric MDS - only makes use of rank order of the dissimilarities. As such, the transformation f need only obey the monotonicity constraint:

$d_{ij} < d_{kl} \Rightarrow f(d_{ij}) \leq f(d_{kl})$

Such an f need only preserve rank order

C. Role of Stress and how it is found

- Stress is the degree of correspondence between the distances among points implied by MDS map and the original matrix input by user is measured by a stress function

Stress defined as $\sum_{i=2}^{n} \sum_{j<i} (\delta_{ij} - d_{ij})^2$

- δ_{ij} is distance between i and j in the plot and d_{ij} is the distance between i and j in the dissimilarity matrix
- want to minimise stress ⇒ higher accuracy

Stress - how well a particular configuration reproduces the observed distance matrix.

D Given Matrix of dissimilarities, explain how classical MDS can be used to lower dimension array X

- Obtain the dissimilarities $\{d_{ij}\}$
- ~~From~~ Form B, each element of which is given by $b_{jk} = -\frac{1}{2}(d^2_{jk} - d^2_{ij} - d^2_{ik})$, with i representing the centroid/origin of all observations
- Create matrix Λ from the eigenvalues $\lambda_1 \dots \lambda_{n-1}$ and the matrix V from the associated eigenvectors $v_1, \dots v_{n-1}$ of B
- Choose an appropriate number of dimensions d, using a suitable measure
- The coordinates of the n required points that are used to represent the n observations in d-dimensional space are given by
  $x_{ij} = \lambda_j^{1/2} v_{ij}$ for $i = 1..n$ $j = 1...d$

E Role of Procrustes Analysis in MDS

- Procrustes Analysis matches one MDS configuration with another by dilation, rotation, reflection and translation
- Aim is to obtain a similar placement and size, by minimising a measure of shape distance call procrustes sum of squares
- ~~Procrustes Sn~~
- Say two MDS methods have been applied to a set of n points resulting in coordinate matrices X and Y
- There is a one to one mapping from i^{th} point in X to i^{th} point in Y

MLA 3

EXAM PAPER 203 MDS Q1

(cont...

The (Procrustes) sum of squared differences between corresponding points in two configurations is: $R^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} (y_{ij} - x_{ij})^2$

- To match configurations, one of them is kept constant (the reference configuration) while the other is translated
- The measure of match between the two configurations is the minimised value of R^2.

F. Relationship of MDS with PCA

- PCA is performed by eigen-decomposition of the data covariance matrix to provide new variables that are formed from linear combinations of the original variables. The new variables are uncorrelated and account for maximum variance in the original variables

- Classical MDS performs eigen decomposition of the data dissimilarity matrix to find a lower-dimensional configuration of the entities such that distances are preserved as closely as possible in a least-squared sense

- When euclidean distance is used within classical MDS, the resulting low dimensional co-ordinates are the same as the principle co ordinates that would be obtained from PCA