MLA

08/05/15. EXAM PAPER 2012 Q1 CLUSTERING.

1 A. We wish to place observations into a group according to their dissimilarity, which has some of the following common properties:
$d(x,y) \geq 0$ 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 $\sqrt{\sum_{i=1}^{n} (x_{in} - x_{jn})^2}$
Manhattan (absolute) $\sum_{i=1}^{n} |x_{in} - x_{jn}|$
Maximum $\max_{k=1,\dots,m} |x_{ik} - x_{jk}|$

In HC, aim to find groups of observations st obs within group are similar and different groups are very dissimilar.

Various methods for measuring dissimilarity between 2 groups A and B:
- 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: $d(A,B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x,y)$

- Complete linkage (can) find clusters of much larger measure of dissimilarity, results in spherical clusters with good internal similarity. Outliers hidden. Invariant under monotonic transform). Small number of large clusters with roughly equal size
- Single - identifies outliers, invariant under monotonic transform, tends towards chaining effect
- Average - result in spherical clusters with good internal similarity, variant under monotonic transformation

B K-Means Clustering
1 Choose number of clusters K and designate cluster centers
2 Assign each data point to the cluster whose center is closest.
3 For cluster i, calculate its centroid $C_i^T = (C(i)_1, C(i)_2 \dots C(i)_m)$ where m is dimension (the number of variables) in an observation (these are found by average value for data point with the cluster)
4 Calculate the sum of squared Euclidean distance of each object to its cluster centroid:
$SS = \sum_{i=1}^{N} \sum_{j=1}^{m} (x_{ij} - C(i)_j)^2$
Assume N observations. Minimize SS
5. Re-assign each observation to the cluster whose centroid is closest
6 Repeat 3-5 until convergence

C - Good for finding circular clusters
- k-means is sensitive to starting points, so running the algorithm several times from different starting values can help check whether results are robust.

D - Looking at HC dendrogram we can see that we generally see that 2 large groups consisting of several smaller sub-groups
- As expected the complete linkage joined the 2 groups at a much larger measure of dissimilarity than average linkage
- Single linkage picked up on an outlier (41) and the chaining effect is particularly noticeable

- For duller analysis, a general rule is to plot K vs. SS and look for a kink in the curve
- If no kink, trade-off between additional complexity by K and better fit by reducing SS
- K=3 is where last significant kink appears in curve

E. Standardisation
- When constructing a dissimilarity matrix for use in clustering, we need to be aware of how our data is scaled
- Important that different variables are comparably scaled, otherwise variable with greater variance will be more prominent in clustering solution
- Some variables are standardised by dividing through by their standard deviation before being used to calculate dissimilarity matrix. Each variable will have variance of 1
- Not always desirable - may want to give less weight to variable with less info
- PCA can be used to reduce dimensionality. Can see the importance of the components via PCA.
- Usually the amount of PCs that account for most variance
- Other method may be used - factor analysis and MDS Cov[Ef]=0

MLA

EXAM PAPER 2012 Q2 FACTOR ANALYSIS

A. Explain Factor Model

- Mathematical approach for attempting to explain the correlation between a large set of variables in terms of a small number of underlying factors
- Main assumption of FA is that it is not possible to observe the underlying factors directly

- Dimensionality of matrix can be reduced from to $m=1$ by expressing the variables as:

$$X_1 = \lambda_1 f + \varepsilon_1$$
$$X_2 = \lambda_2 f + \varepsilon_2$$
$$X_m = \lambda_m f + \varepsilon_m$$

- The f in these equations is an underlying common factor, the λ_i's are known as factor loadings, whilst the ε_i's are known as errors or specific factors
- Common factor can often be given a general interpretation like "general ability"
- The specific factors ε_i will have small variance if X_i is closely related to general ability

- The observable random vector $X^T = (X_1, X_2, \ldots X_m)$ had mean μ and covariance matrix Σ
- Factor Model stated that X is linearly independent upon a few unobservable random variables $f_1, f_2 \ldots f_p$ called common factors and m additional sources of variation $\varepsilon_1 \ldots \varepsilon_m$ called Specific factors

$$X_1 - \mu_1 = \lambda_{11} f_1 + \lambda_{12} f_2 \ldots \lambda_{1p} f_p + \varepsilon_1$$
~~$X_2 - \mu_2 = X_2$~~
$$X_m - \mu_m = \lambda_{m1} f_1 + \lambda_{m2} f_2 \ldots + \lambda_{mp} f_p + \varepsilon_m$$

- The λ_{ij} value is called the factor loading of i^{th} variable on j^{th} factor
- Λ is matrix of Factor loadings
- Note that i^{th} Specific factor ε_i is associated only with response X_i
- $f_1 \ldots f_m$ and $\varepsilon_1 \ldots \varepsilon_m$ are all unobservable random variables

MA4

04/05/18 EXAM PAPER 202 Q2 FACTOR ANALYSIS OTHER METHODS

Q2 A Factor Model

- An observable random vector $X = (X_1 \dots X_m)$ has mean μ and covariance matrix Σ
- Factor model states that X is linearly dependent upon a few unobservable random variables $F_1 \dots F_p$ called common factors and m additional sources of variation $\varepsilon_1 \dots \varepsilon_m$ called specific factors

Here: $X_1 - \mu_1 = \lambda_{11} F_1 + \lambda_{12} F_2 + \dots \lambda_{1p} F_p + \varepsilon_1$

$X_m - \mu_m = \lambda_{m1} F_1 + \lambda_{m2} F_2 + \dots \lambda_{mp} F_p + \varepsilon_m$

- where λ_{ij} is called the factor loading of the i^{th} variable on the j^{th} factor
- Under the model, it is assumed that

$E[F] = 0$ $\quad Cov[F] = I = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$

$E[\varepsilon] = 0$ $\quad Cov[\varepsilon] = \Psi = \begin{pmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_m \end{pmatrix}$

$Cov[\varepsilon, F] = 0$

B Variance - Communality and Uniqueness

- The variance of a MV random variable X can be split into 2 parts.
- First portion of the variance for the i^{th} component arises from the m common factors and is referred to as the communality
- Remainder of the variance for i^{th} component is due to the specific factor, and is referred to as the uniqueness.
- Denoting i^{th} communality by h_i^2:

$\sigma_{ii}^2 = \lambda_{i1}^2 + \lambda_{i2}^2 \dots + \lambda_{ip}^2 + \psi_i$ — Sum of squared loadings of i^{th} variable on p common factors

$= h_i^2 + \psi_i$

i.e. Var[x_i] = communality + uniqueness

C Factor Rotation and Varimax Rotation

- If initial loadings are subject to an orthogonal transformation (i.e. multiplied by an orthogonal matrix G), the covariance matrix can still be reproduced
- An orthogonal transformation corresponds to a rigid rotation or reflection of coordinate axes.
- Thus, an orthogonal transformation of the factor loadings (and implied transformation of factors) is called a factor rotation.
- Irrelevant mathematically, but when the objective is statistical interpretation, may be that one rotation is more useful than another.

- The varimax procedure selects the orthogonal transformation G maximising V, the sum of the column variances) across all factors
- Maximising V corresponds to 'spreading out' the squared of the loadings on each factor as much as possible
- Hence groups of large and negligible coefs are found in any column of Λ^* i.e. it aims for a rotation that makes the squared loadings λ^2 either large or small i.e. few medium sized coefs

DAVID WETHERELL

D Output interpretation
- Communalities = Sum of squared loadings in each row of loading matrix
- Communalities + Uniquenesses for each variable = total variance for that variable
SS-loadings = Sum of squared loadings for their factor.
- Proportion variance = SS/m = SS/6 in this case
- The two factors account for 59.7% of total variance

- Factor 1: literacy: 38% variance. Correlated to good reading abilities
- Factor 2: Spatial awareness. Focused on visual factors

- Reading seems to be best explained by small uniqueness - smaller the uniqueness, larger the variance in the model

E PCA, MDS, and FA
- PCA finds linear combinations of the variables in the data which capture most of the variation in the original data
- MDS seeks to produce lower dimensional summary of data, such that distances between i and j are in the representation d_{ij} are close to dissimilarities between their points or d_{ij} if (i,j).
- Takes a set of dissimilarities and returns a set of dimensional points s.t. the distances between points are dissimilarities
- Classical MDS performs eigen decomposition of data dissimilarity matrix to find low-dimensional such that distances are preserved as closely as possible in a least squares sense
- FA shares same objective as PCA but is much more elaborate
- Not affected by rescaling unlike PCA.
- PCA, MDS no data assumptions FA assumed data came from multivariate normal distribution in which assumption held i.e. $E[F]=0$
- PCA: data $\Rightarrow$ PCs FA: factors $\Rightarrow$ data
- PCA - when specific variables are lost, often distorted into PCs. Interpret FA model special problem for them. When small PCA and FA produce similar results

- Assumed that:

$E[F] = 0$ $\quad Cov[F] = I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$E[\varepsilon] = 0$ $\quad Cov[\varepsilon] = \Psi = \begin{pmatrix} \psi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \psi_m \end{pmatrix}$

F and ε are independent so $Cov[F, \varepsilon] = 0$

B. Variance, Communality and uniqueness

- Variance of X can be split into 2 parts
- The first portion of variance for i^{th} component arises from the m common factors and is referred to as the i^{th} communality
- Remainder of variance for i^{th} component is due to Specific factor, referred to as uniqueness
- Denoting the i^{th} communality as h_i^2 then:

$$\sigma_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots \lambda_{ip}^2 + \psi_i$$
$$= h_i^2 + \psi_i^2$$

$Var[x_i]$ = communality + uniqueness

- i^{th} communality is the sum of squares of the loadings of i^{th} variable on p common factors

C. Factor Rotation and Varimax Rotation

- If initial loadings are subject to an orthogonal transformation (i.e. multiplied by an orthogonal matrix G), the covariance matrix Σ can still be reproduced
- An orthogonal transformation corresponds to a rigid rotation or reflection of coordinate axis
- Hence, the orthogonal transformation of factor loadings (and the implied transformation of the factors) is called a factor rotation
- VARIMAX - Note that the squared loadings λ_{ij}^2 is the proportion of the variance that is attributable to common factor j:

$Var[x_i] = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots \lambda_{ip}^2 + \psi_i \quad = h_i^2 + \psi_i$

- We aim for a rotation which makes the loading either large or small (i.e. far from medium sized values)
- Varimax procedure selects the orthogonal transformation G maximising the sum of column variances over all factors $j = 1 \dots p$

MLA

EXAM PAPER Q2 FACTOR ANALYSIS

Meaning V corresponds to "Speaking at" the square of the loadings on each factor as much as possible. Hence, both groups of large and negligible coefficients are found in any column of Λ^*

D) Interpretation of output.

- Communalities are the sum of square loadings in each row of loading matrix
- The communality for each variable added to the uniqueness for each variable is the total variance for that variable. (equal to 1 if data is standardised)
- The SS loadings are the sum of squared loadings for each factor (column).
- Cumulative = 59.7% of variance (of data) is accounted for by 2 factors

Factor 1: Higher loading on reading, words and general, this could be interpreted as English skill/knowledge factor

Factor: High loading on block, picture and general and maze, could be interpreted as visualisation more mind factor or problem solving factor.

E. Other Dimension Reduction Method

PCA v FA

- PCA looks for linear combination of the data matrix X that are uncorrelated and of high variance, whilst FA seeks uncorrelated linear combinations of the variables responsible underlying fundamental quantities
- PCA makes no assumptions about form of covariance matrix, whilst FA assumes data comes from a well defined model in which specific assumptions hold e.g. $\mathbb{E}[\varepsilon]=0$, $\mathbb{I}[\varepsilon]=0$
- PCA: data $\Rightarrow$ PCs. FA: factors $\Rightarrow$ data
- When specific variances are large they are absorbed into the PCs whereas FA would spread them. When specific variances are small, PCA and FA give similar results
- Two analysis often performed together. Example $\rightarrow$ can conduct a PCA to determine number of factors to extract in FA

FA v MDS

- FA requires that the underlying data are distributed as Multivariate Normal
that the relationships are linear. MDS imposes no such restrictions
- As long as rank ordering of distances (or similarities) in matrix is meaningful, MDS can be used

- In terms of resultant differences, FA tends to extract more factors (dimensions)
than MDS, as a result MDS often yields more readily interpretable solutions
- MDS can be applied to any kind of distances or similarities, while FA requires
us to first compute a correlation matrix

- MDS can be based on subjects' direct assessment of similarity between stimuli, while
FA requires subjects to rate those stimuli on some list of attributes

MLA

EXAM PAPER 2012 Q1

1 A Dissimilarity Measure

Euclidean $\sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$

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

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

LINKAGE

Single: $d(A,B) = \min_{x \in A} \min_{y \in B} d(x,y)$ result in spherical cluster, good internal symmetry, displays outlier, invariant under monotone transformation

Complete: $d(A,B) = \max_{x \in A} \max_{y \in B} d(x,y)$ Join final clusters at much larger method of dissimilarity, result in spherical clusters, good internal symmetry, highly outlier, not invariant under monotone transformation, small number of large clusters

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

variant under monotone transform

B K-Means Clustering Algorithm

1. Choose number of clusters k and designate cluster centers.
2. Assign each data point to the cluster whose center is closest.
3. For cluster i, calculate the centroid $C_i^T = (C(i)_1, C(i)_2 \ldots C(i)_m)$ where m denotes the number of variables in an observation.
4. Calculate the sum of squared distance of each object to its cluster centroid:
   $SS = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - C(i)_j)^2$
   Assume total of n observations want to minimize SS
5. Re-assign each observation to the cluster whose centroid is closest.
6. Repeat (3)-(5) until convergence

C Different start points may give different results want to find convergence

K-means clustering work for circular shapes

D 2-groups From Scree plot and graph

E Standardization prevent variable with larger variance appears prominently in the clustering result.
Gives extra weight to small variables also

Dimension reduction will make a easier solution more interpretable, will remove non-significant variables from analysis