Mining big data by multivariate methods

Luigi Palla- Assistant Prof. in Medical Statistics and Epidemiology

Medical Statistics Dept

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Centre for Statistical Methodology theme

Multivariate Statistical methods comprise:

- **Methods based on a statistical model** (i.e. with a model having a probability distribution) like:

-factor analysis; linear discriminant analysis; partial least squares; reduced rank regression ect…

- **Machine learning methods** like:

-classification trees; random forests; support vector machines; neural networks…

- **Descriptive (mathematical/geometric) methods** like:

-principal component analysis; multidimensional scaling; cluster analysis; correspondence analysis; Procrustes Analysis.

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Descriptive multivariate statistical methods

- Methods to analyse data matrices , for example:

-p variables (matrix columns) for n statistical units (matrix rows);

-matrix of variances/covariances between p variables;

-matrix of distances/dissimilarities between n statistical units;

-contingency table cross-classifying two categorical variables.

- Underpinned by matrix algebra and geometry rather than probabilistic statistical theory.

- Allow graphical exploration of the data: important tool for big data analysis!

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Mining Data by Correspondence Analysis (CA) in a nutrition survey

(joint work with Andrew Chapman)

- National Diet and Nutrition Survey Rolling Programme data (2008-2012) contains 4 day diet diary for representative sample of the British population.
- Contains also information about where food is consumed.

AIM: Investigation on the relationship between type of foods and location in adolescents.

- Context of eating is a modifiable exposure.
- 884 teenagers provide information on 62,523 eating occasions

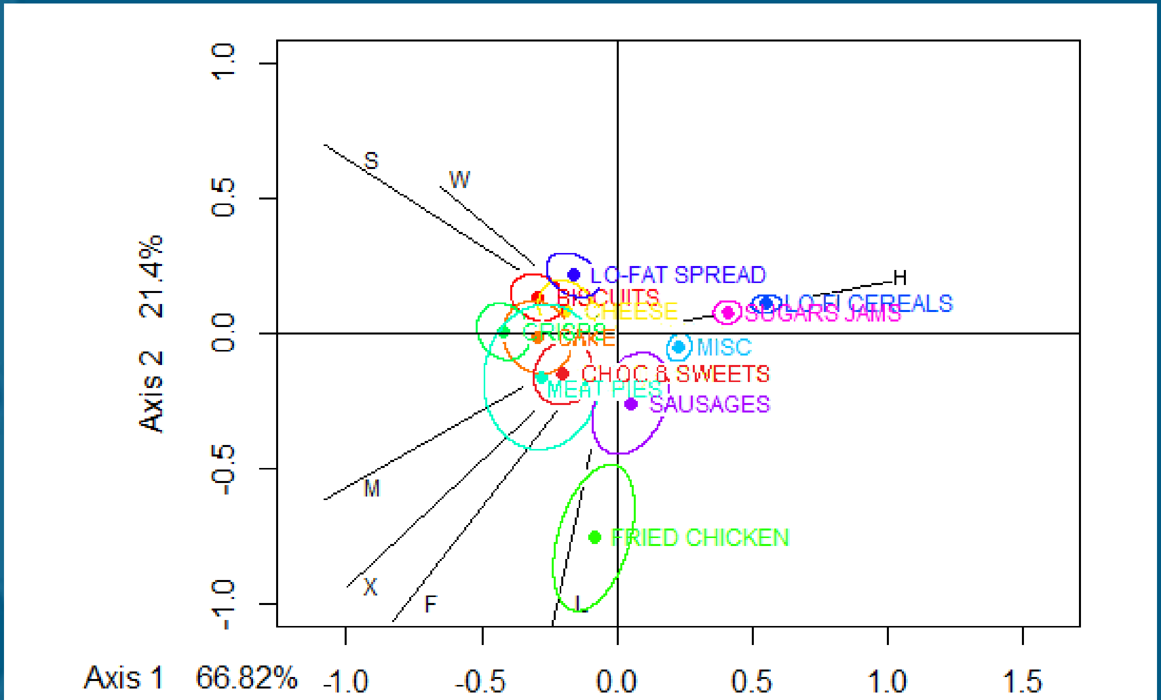
LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Hypothesis generation by CA

- More than a hundred food categories are reduced to 25 based on their cumulative contribution to total energy intake.
- They are also categorised as Healthy, Neutral, Less Healthy according to their nutrient composition.
- Half of the eating occasions are randomly sampled for hypothesis generation.
- The reduced set of food categories are cross-classified by the eating location producing a contingency table (data matrix).
- CA produces a graphical representation of the association between foods and location via a biplot.

Biplot of CA of less-healthy foods and locations

- 87% of variation in the data matrix is represented
- 95% bootstrap Confidence Regions for Food Groups account for sample variability.



None of the ellipses includes the origin so they differ significantly from the average profile.
Legend: H-Home S-School W-Work F-Friends/Carers L-Leisure M-Mobile X-Other

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Hypothesis testing via Logistic Regression (GEE)

Food-Group	OR Other vs Home	99% CI p-value	OR Other vs School-Work	99% CI p-value
Chocol. and Sweets	2.5	(1.8, 3.4) p<0.0001	1.8	(1.2, 2.8) p=0.0002
Meat Pies	2.8	(1.5, 5.0) p<0.0001	1.3	(0.6, 3.0) p=0.44

Table 1. Adjusted Odds Ratio estimates for food-groups as outcomes and location-types as exposures.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Conclusions

- CA allowed to mine complex survey data for potential associations between outcome and exposure.
- The hypotheses generated via such geometric method avoid multiple testing issue and can then be tested formally by regression analysis for dependent data.

- In general multivariate descriptive methods can complement traditional analysis method in medical statistics.
- Further research and efforts (teaching/seminars) are warranted in order to promote multivariate methods use and integration for big data analysis in epidemiology.

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE