A Comparison between Latent Semantic Analysis and Correspondence Analysis

Julie Séguéla[1,2] **, Gilbert Saporta**[1,*]
1. CEDRIC, CNAM, 292 rue Saint-Martin, F-75141 Paris cedex 03
2. Multiposting.fr, 33 rue Réaumur, 75003 Paris
* Contact author: gilbert.saporta@cnam.fr

Keywords: Latent semantic analysis, textual data, correspondence analysis, web data

Latent Semantic Analysis (LSA) is a technique for analyzing textual data through a singular value decomposition of term-document matrices (Deerwester et al. (1990), Landauer et al. (2007)). The basic postulate is that there is an underlying latent semantic structure in word usage data that is partially hidden or obscured by the variability of word choice (synonymy problem). LSA is also called Latent Semantic Indexing (LSI) in information retrieval, where the main application consists in computing similarities between user's query and all documents in the space, or between documents.

Since LSA is a SVD of a contingency table, it strongly resembles to Correspondence Analysis (CA), see Lebart et al. (1998). Before performing the SVD, practitioners of LSA recommend several weighting functions of the frequencies, but not the one leading to the chi-square metric. Typically, LSA allows to reduce the dimensionality from several thousands to several hundred of a huge but sparse data matrix. Given the dimension, graphical representations are useless. In the context of statistical implementations, the coordinates can be used for categorization tasks (in supervised or unsupervised frameworks).

We first compare basic LSA with CA on a toy example. Then performances of CA and LSA with several weighting functions are compared on a large data set coming from job offers posted on the web. When posted on the internet, job offers have been labeled by recruiters according to the job category (e.g. Marketing, Information Systems, Finance, etc.). We are interested in the capacity of these document representation technics to lead us to the real job category with a clustering method. After preprocessing of job offers, we compute similarities between texts based on coordinates in reduced spaces and apply an hybrid method combining hierarchical clustering and k-means algorithm. Performance of text representation methods will be assessed with three different measures (Cohen's Kappa, Rand index, F-measure) and discussed according to the number of dimensions kept.

References

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science*, **41**, 391-407.

Landauer, T. K., & al. (2007). *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates.

Lebart, L., Salem, A. & Berry, L. (1998). *Exploring Textual Data*, Kluwer.

LSA website,
`http://lsa.colorado.edu/`.