The correspondence analysis of partitioned tables with multiple factors

Koen Plevoets, Isabelle Delaere & Gert De Sutter

University College Ghent / Ghent University – Faculty of Applied Language Studies
Groot-Brittanniëlaan 45
B-9000 Ghent

This paper presents a modification of Correspondence Analysis (Greenacre 2007) which is customised to analysing partitioned data in relation to multiple explaining factors. A linguistic variable is typically represented as a categorical variable (Labov 1966); however, traditional linguistic inquiries have been restricted to the study of single linguistic variables. In contrast, there is a growing interest to broaden the scope to several variables and foray into the analysis of linguistic "varieties", as is the objective in *stylometry* (Biber 1995) and *sociolectometry* (Geeraerts et al. 1999, Speelman et al. 2003).

Our technique taps into these approaches as, firstly, it applies a partitioning of the linguistic data into sets of various synonyms. Secondly, the technique does not merely analyse the correlations between variables (as is customary in Multiple Correspondence Analysis; Greenacre & Blasius 2006), but cross-tabulates the linguistic variants on the one hand with the combination of all explanatory factors on the other; this enables the study of interactions between the factors. Finally, statistical inference is implemented by means of the bootstrap procedure developed for Correspondence Analysis (Lebart et al. 2003). By consequence, our technique seeks a middle ground between regression-like techniques such as Loglinear Analysis (Agresti 2002) on the one hand and the framework of *Geometric Data Analysis* (Le Roux & Rouanet 2010) on the other, where high-dimensional data are mapped in a reduced space (which is also characteristic of text mining techniques such as Latent Semantic Analysis; Landauer & Dumais 1997).

The technique will be illustrated by a case study involving translational differences with respect to various text genres and the effect of source language (Delaere et al. accepted, De Sutter et al. 2012). The results show a. o. that there is a distinction between well-edited genres and genres with less editorial control, and that translations are overall more normalised than non-translations.

Bibliography

Agresti, A. (2002). *Categorical data analysis*. Hoboken: Wiley.

Biber, D. (1995). *Dimensions of register variation. A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Delaere, I., G. De Sutter & K. Plevoets (accepted). "Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties". *Target*.

De Sutter, G., I. Delaere & K. Plevoets (2012). "Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modeling". In: Oakes, Michael & Meng Ji (eds), *Quantitative Methods in Translation Studies*. Amsterdam/Philadelphia: John Benjamins, 325-345.

Geeraerts, D., S. Grondelaers & D. Speelman (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut.

Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton: Chapman and Hall/CRC.

Greenacre, M. & J. Blasius (2006). *Multiple correspondence analysis and related methods*. Boca Raton: Chapman and Hall/CRC.

Labov, W. (1966). "The linguistic variable as a structural unit". *Washington Linguistics Review* 3, 4-22.

Landauer, T. K. & Dumais, S. T. (1997). "A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and the representation of knowledge". *Psychological Review, 104*, 211-240.

Lebart, L., M. Piron & J.-F. Steiner (2003). *La sémiométrie*. Paris: Dunod.

Le Roux, B. & H. Rouanet (2010). *Geometric data analysis. From correspondence analysis to structured data analysis*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Speelman, D., S. Grondelaers & D. Geeraerts (2003). "Profile-based linguistic uniformity as a generic method for comparing language varieties". *Computers and the Humanities* 37, 317-337.