Combining PCA and GLM to overcome data sparsity in the multivariate analysis of linguistic variation

Mahdi Mantash & Stephanie Evert
Friedrich-Alexander-Universität Erlangen-Nürnberg
[mahdi.mantash, stephanie.evert]@fau.de

Multivariate statistical analysis with latent variable models is a highly successful approach to studying linguistic variation, where it is applied to per-text vectors of lexico-grammatical and text-statistical features. Its fundamental assumption is that correlational patterns between such features reflect underlying dimensions of variation that can be identified by latent variable models. Linguistic interpretation is based on the feature weights contributing to each dimension and on visualisation of the latent coordinates of the individual texts (e.g. Neumann & Evert 2021). Two popular statistical techniques are factor analysis (FA), which is central to multidimensional analysis (MDA: Biber 1988), and principal component analysis (PCA), which plays a similar role for geometric multivariate analysis (GMA: Neumann & Evert 2021). Despite mathematical differences, both methods often produce very similar results in linguistic studies.

FA and PCA share a weakness that limits their applicability to short texts: they assume Gaussian distributions of the feature vectors. Very short texts (less than a few hundred words) and sparse features (such as imperatives) violate this assumption because they inflate sampling variation and lead to skewed distributions. The resulting outlier data points make the identification of latent dimensions unreliable and affect the interpretability of visualisations.

In this contribution, we address this problem by combining probabilistic PCA (Tipping & Bishop 1999) with generalised linear models (GLM: McCullagh & Nelder 1989), using their link function and binomial sampling distribution to account explicitly for the properties of short texts and sparse features. The crucial advantage of our approach is its ability to quantify uncertainty in the observed data, giving shorter texts less influence on the latent dimensions and providing posterior distributions for their latent coordinates. In a case study based on data from the International Corpus of English, we compare the new method to traditional PCA and FA, looking both at register variation and differences between varieties.

References: • Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge. • McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London, New York: Chapman and Hall. • Neumann, S. & S. Evert (2021). A register variation perspective on varieties of English. In E. Seoane & D. Biber (eds.), *Corpus-Based Approaches to Register Variation*, Amsterdam: Benjamins, 143–178. • Tipping, M. E. & C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61(3), 611–622.