Towards Making sense of multivariate analyses of linguistic variation

Stefan Evert

Multidimensional analysis (Biber 1988)
- 481 texts, 67 lexicogrammatical features
- unsupervised FA
- validation: separation of “known” genre categories

Problems
- choice of features & texts
- interpretation of FA weights

Case study I: Authorship attribution with Burrows’s Delta
unsupervised clustering
- 25 authors x 3 novels for EN, DE, FR
- 200 – 5000 features
- Ward clustering / PAM

$$\Delta_B(D_1, D_2) = \sum_{i=1}^{n} |z_i(D_1) - z_i(D_2)|$$

Problems
- only 75 texts
- how & why does $$\Delta_B$$ work so well?

Case study II: Evidence for shining-through in translations
minimally supervised PCA (linear discriminant analysis)
- 298 texts from CroCo corpus
  (78x EN→DE, 71x DE→EN)
- 27 features grounded in SFL

Problems
- interpretation of LDA weights
- are weights stable or do they depend on choice of texts?
- is our selection of features crucial to the results?

Interpretation of dimension weights
- standard approach based on magnitude and sign of weights
  (EN on positive side of axis)
- interprets features as correlated rather than complementary

- better approach: what does each feature contribute to the LDA positions of texts?
- reveals entirely different patterns
- correlated features help LDA to reduce within-group variance

What are the characteristic words?
- supervised recursive feature elimination
  → 233 words as features
- not just mfw, but none unique to one author
- with, so, i, But, And, upon, don, head,
  Then, looking, almost, indeed, nor, …,
  XXXVIII (df=34), XLII (df=29), XLIII (df=26),
  hereabout (df=11), viest (df=15), contours (df=9), Ecod (df=4), …
- validation for DE: new novels from same authors: 97% accuracy

Work in progress
- contribution of features to silhouette width of clustering
- assess relevance to each author
- identify features responsible for mis-classifications

Bootstrapping latent dimensions
- bootstrapping / cross-validation can be used to assess stability of LDA &PCA dimensions (algorithmic because of small # of features)
- LDA axis “wobbles” by approx. 10° across folds
- moderate variability of feature weights: $$\alpha < 0.05$$
- but positions of texts on LDA axis are stable ($$r = .987$$)

Reliability of the clustering
- bootstrapping texts not applicable to clustering & high-dimen. feature space
- bootstrapping features → unclear
- biggest factor: choice of authors (empirical study on Gutenberg archive)

www.linguistik.fau.de | www.stefan-evert.de