Unit 7: Multivariate Analysis

Statistics for Linguists with R – A SIGIL Course

Designed by Stefan Evert[1] and Marco Baroni[2]

[1]Computational Corpus Linguistics Group
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

[2]Center for Mind/Brain Sciences (CIMeC)
University of Trento, Italy

Outline

Outline

What is multivariate analysis?

- Univariate statistics
  - focus on a single variable of interest (at a time)
  - estimate population parameters (π, μ, σ^2, ...)
  - comparison of two or more groups
- Bivariate statistics
  - focus on interdependencies of two variables
  - correlation & co-occurrence
- Regression modelling
  - predict single target variable ("dependent")
  - based on multiple other variables ("independent")
- Multivariate statistics
  - combined effects of many variables
  - correlations & distribution patterns
  - often "unsupervised": no target variable or comparison groups

Application examples

- Register variation (Biber 1988, 1993)
- Translation studies (Evert & Neumann 2017; De Sutter *et al.* 2012)
- Stylometry: authorshop attribution (Evert *et al.* 2017)
- Dialectology (Speelman *et al.* 2003)
- Historical linguistics (Sagi *et al.* 2009; Perek 2018)
- Identification of confounding variables (Tummers *et al.* 2014)
- Linguistic productivity (Jenset & McGillivray 2012)
- Correspondence analysis (Greenacre 2007)
- Distributional semantics (see ESSLLI course)

Outline

R packages

Required R packages:

- `corpora` ($\geq$ 0.5)
- `wordspace` ($\geq$ 0.2)

Recommended packages:

- `ggplot2`, `reshape2` ... for plotting feature weights
- `rgl` ... for interactive 3-d visualization
- `Hotelling`, `ellipse` ... for significance testing
- `e1071` ... for machine learning (SVM)
- `Rtsne` ... for low-dimensional maps
- `ca` ... for correspondence analysis

☞ install with package manager in RStudio or R GUI

Code & data sets

Download additional code & data sets from SIGIL homepage:

- `multivar_utils.R`
- `unit7_data.rda`

☞ put all files in RStudio project directory (or working directory)

```
> library(corpora)              # basic utilities and some data sets
> library(wordspace)            # for large and sparse matrices

> source("multivar_utils.R")    # additional functions

> load("unit7_data.rda", verbose=TRUE)  # further data sets
```

Overview of data sets

- 65 Biber features for British National Corpus
  - `BNCbiber` = 4048 × 65 feature matrix
  - `BNCmeta` = complete metadata table
  - extensive documentation with `?BNCbiber`, `?BNCmeta`
- 67 Biber features for Brown Family corpora
  - `BrownBiber_Matrix` = 3500x67 feature matrix
  - `BrownBiber_Meta` = metadata table
  - features are Biber-scaled z-scores obtained with MAT v1.3
    http://sites.google.com/site/multidimensionaltagger/
  - see tagger manual for feature definitions

Overview of data sets

- 27 SFL-inspired features for translation pairs (CroCo corpus)
  - `CroCo_Matrix` = 452 × 27 feature matrix
  - `CroCo_Meta` = metadata table
  - `CroCo_orig2trans` = row numbers of translation pairs
  - data from Evert & Neumann (2017)
- Literary authorship attribution with Δ measures
  - data: sparse document-term matrices for 20,000 most frequent words (mfw) as `wordspace` DSM objects
  - `Delta$DE` = 75 × 20000 matrix (German novels, 25 authors)
  - `Delta$EN` = 75 × 20000 matrix (English novels, 25 authors)
  - `Delta$FR` = 75 × 20000 matrix (French novels, 25 authors)
  - `Delta$DE$rows`, `Delta$EN$rows`, ... = metadata tables
  - `DeltaLemma` = lemmatized version
  - data from Jannidis *et al.* (2015); Evert *et al.* (2017)

Overview of data sets

- 19 type-token complexity measures for Δ corpus
  - complexity scores for 10,000-token text slices from 75 novels
  - `DeltaComplexity$DE$Matrix` = 996 × 19 matrix (German)
  - `DeltaComplexity$EN$Matrix` = 1147 × 19 matrix (English)
  - `DeltaComplexity$FR$Matrix` = 679 × 19 matrix (French)
  - `DeltaComplexity$DE$Meta`, ... = metadata tables
  - can be used to study correlational patterns between measures
- 7 syntactic complexity measures for 969 German novels
  - `SyntacticComplexity_Matrix` = 969 × 7 feature matrix
  - `SyntacticComplexity_Meta` = metadata tables
  - can be used to compare high-brow against low-brow literature

Outline

Feature matrix

Feature matrix records quantitative features for each text

$$\mathbf{M} = \begin{bmatrix} \cdots & \mathbf{m}_1 & \cdots \\ \cdots & \mathbf{m}_2 & \cdots \\ & \vdots & \\ & \vdots & \\ \cdots & \mathbf{m}_k & \cdots \end{bmatrix}$$

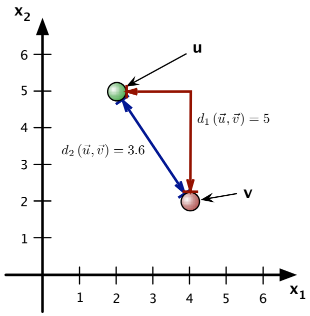
	nominal	pass	prep	subord	ttr
$orig_1$	1.205	5.013	6.883	4.483	1.285
$orig_2$	0.738	2.537	6.486	6.157	1.714
$orig_3$	1.252	4.462	8.463	4.785	2.476
$orig_4$	1.105	2.899	8.119	3.966	1.519
$orig_5$	1.764	4.268	7.167	3.947	1.792
$orig_8$	1.545	7.268	7.461	5.455	1.572
$trans_1$	0.463	2.208	6.297	6.089	2.339
$trans_2$	1.131	2.597	6.307	4.844	1.810
$trans_4$	0.935	1.744	7.098	4.012	1.403
$trans_5$	0.867	3.604	7.511	5.154	1.902
$trans_7$	1.387	4.290	8.211	3.998	1.822

```
> M <- MultiVar_Matrix
> M
```


Outline

Geometric distance = metric

- **Distance** between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ ➜ (dis)similarity
  - $\mathbf{u} = (u_1, \ldots, u_n)$
  - $\mathbf{v} = (v_1, \ldots, v_n)$
- **Euclidean** distance $d_2(\mathbf{u}, \mathbf{v})$
- "City block" **Manhattan** distance $d_1(\mathbf{u}, \mathbf{v})$
- Both are special cases of the **Minkowski** p-distance $d_p(\mathbf{u}, \mathbf{v})$ (for $p \in [1, \infty]$)

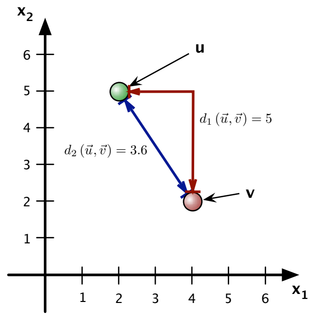


$$d_p(\mathbf{u}, \mathbf{v}) := \left(|u_1 - v_1|^p + \cdots + |u_n - v_n|^p\right)^{1/p}$$

$$d_\infty(\mathbf{u}, \mathbf{v}) = \max\left\{|u_1 - v_1|, \ldots, |u_n - v_n|\right\}$$

Geometric distance = metric

- **Distance** between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ ➜ (dis)similarity
  - $\mathbf{u} = (u_1, \ldots, u_n)$
  - $\mathbf{v} = (v_1, \ldots, v_n)$
- **Euclidean** distance $d_2(\mathbf{u}, \mathbf{v})$
- "City block" **Manhattan** distance $d_1(\mathbf{u}, \mathbf{v})$
- Extension of p-distance $d_p(\mathbf{u}, \mathbf{v})$ (for $0 \le p \le 1$)

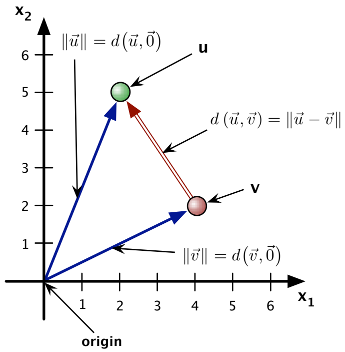


$$d_p(\mathbf{u}, \mathbf{v}) := |u_1 - v_1|^p + \cdots + |u_n - v_n|^p$$

$$d_0(\mathbf{u}, \mathbf{v}) = \#\{i \mid u_i \neq v_i\}$$

Distance and vector length = norm

- Intuitively, distance $d(\mathbf{u}, \mathbf{v})$ should correspond to length $\|\mathbf{u} - \mathbf{v}\|$ of displacement vector $\mathbf{u} - \mathbf{v}$
  - $d(\mathbf{u}, \mathbf{v})$ is a **metric**
  - $\|\mathbf{u} - \mathbf{v}\|$ is a **norm**
  - $\|\mathbf{u}\| = d(\mathbf{u}, \mathbf{0})$
- Any norm-induced metric is **translation-invariant**



- $d_p(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_p$
- **Minkowski p-norm** for $p \in [1, \infty]$ (not $p < 1$):

$$\|\mathbf{u}\|_p := \left(|u_1|^p + \cdots + |u_n|^p\right)^{1/p}$$

Computing distances

Compute distances between all pairs of texts:

```
> round(dist(M), 2)   # returns a triangular 'dist' object
> round(dist(M, method="manhattan"), 2) # Manhattan metric
```

Use `wordspace` function for additional metrics:

```
> dist.matrix(M, method="mink", p=0.5)   # full matrix
> dist.matrix(M, method="mink", p=0.5, as.dist=TRUE)
```

Standardize features for equal contribution to Euclidean metric:

```
> Z <- scale(M)       # matrix of z-scores
> round(dist(Z), 2)  # default: Euclidean metric
```


Outline

Linear subspace & basis

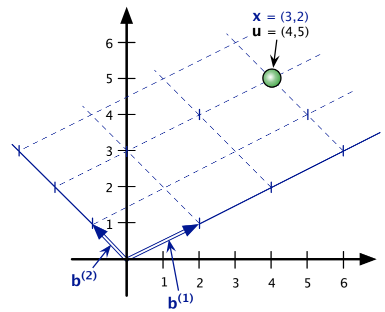
- A linear **subspace** $B \subseteq \mathbb{R}^n$ of rank $r \leq n$ is spanned by a set of r linearly independent basis vectors

$$B = \{\mathbf{b}_1, \ldots, \mathbf{b}_r\}$$

- Every point $\mathbf{u}$ in the subspace is a unique linear combination of the basis vectors

$$\mathbf{u} = x_1 \mathbf{b}_1 + \ldots + x_r \mathbf{b}_r$$

- Coordinate vector $\mathbf{x} \in \mathbb{R}^r$ with respect to the basis



Linear subspace & basis

- Basis matrix $\mathbf{V} \in \mathbb{R}^{n \times r}$ with column vectors $\mathbf{b}_i$:

$$\mathbf{u} = x_1 \mathbf{b}_1 + \ldots + x_r \mathbf{b}_r = \mathbf{V}\mathbf{x}$$

$$\begin{bmatrix} x_1 b_{11} + \ldots + x_r b_{1r} \\ x_1 b_{21} + \ldots + x_r b_{2r} \\ \vdots \\ x_1 b_{n1} + \ldots + x_r b_{nr} \end{bmatrix} = \begin{bmatrix} b_{11} & \cdots & b_{1r} \\ b_{21} & \cdots & b_{2r} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nr} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_r \end{bmatrix}$$

$$\underset{(n \times 1)}{\mathbf{u}} = \underset{(n \times r)}{\mathbf{V}} \cdot \underset{(r \times 1)}{\mathbf{x}}$$

An aside: Matrix multiplication

$$\begin{bmatrix} & & \\ a_{ij} & & \\ & & \end{bmatrix} = \begin{bmatrix} & & \\ b_{i1} & \cdots & b_{in} \\ & & \end{bmatrix} \cdot \begin{bmatrix} & c_{1j} & \\ & \vdots & \\ & \vdots & \\ & c_{nj} & \end{bmatrix}$$

$$\underset{(k \times m)}{\mathbf{A}} = \underset{(k \times n)}{\mathbf{B}} \cdot \underset{(n \times m)}{\mathbf{C}}$$

- **B** and **C** must be **conformable** (in dimension n)
- Element a_{ij} is the inner product of the i-th row of **B** and the j-th column of **C**

$$a_{ij} = b_{i1} c_{1j} + \ldots + b_{in} c_{nj} = \sum_{t=1}^{n} b_{it} c_{tj}$$

Orthonormal basis

- Particularly convenient with orthonormal basis:

$$\|\mathbf{b}_i\|_2 = 1$$
$$\mathbf{b}_i^T \mathbf{b}_j = 0 \qquad \text{for } i \neq j$$

- Corresponding basis matrix **V** is (column)-**orthogonal**

$$\mathbf{V}^T \mathbf{V} = \mathbf{I}_r$$

and defines a Cartesian coordinate system in the subspace

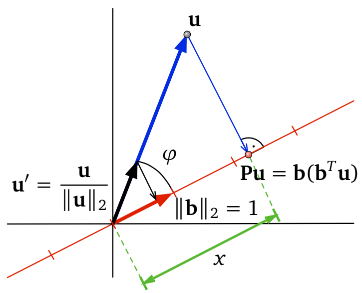
☞ From now on always assume orthonormal basis

The mathematics of projections

- 1-d subspace spanned by basis vector $\|\mathbf{b}\|_2 = 1$
- For any point $\mathbf{u}$, we have

$$\cos\varphi = \frac{\mathbf{b}^T \mathbf{u}}{\|\mathbf{b}\|_2 \cdot \|\mathbf{u}\|_2} = \frac{\mathbf{b}^T \mathbf{u}}{\|\mathbf{u}\|_2}$$

- Trigonometry: coordinate of point on the line is $x = \|\mathbf{u}\|_2 \cdot \cos\varphi = \mathbf{b}^T \mathbf{u}$



- The projected point in original space is then given by

$$\mathbf{b} \cdot x = \mathbf{b}(\mathbf{b}^T \mathbf{u}) = (\mathbf{b}\mathbf{b}^T)\mathbf{u} = \mathbf{P}\mathbf{u}$$

where **P** is a **projection matrix** of rank 1

The mathematics of projections

- For an orthogonal basis matrix $\mathbf{V}$ with columns $\mathbf{b}_1, \ldots, \mathbf{b}_r$, the projection into the rank-r subspace B is given by

$$\mathbf{P}\mathbf{u} = \left(\sum_{i=1}^{r} \mathbf{b}_i \mathbf{b}_i^T \right) \mathbf{u} = \mathbf{V}\mathbf{V}^T \mathbf{u}$$

  and its subspace coordinates are $\mathbf{x} = \mathbf{V}^T \mathbf{u}$
- Projection can be seen as decomposition into the projected vector and its orthogonal complement

$$\mathbf{u} = \mathbf{P}\mathbf{u} + (\mathbf{u} - \mathbf{P}\mathbf{u}) = \mathbf{P}\mathbf{u} + (\mathbf{I} - \mathbf{P})\mathbf{u} = \mathbf{P}\mathbf{u} + \mathbf{Q}\mathbf{u}$$

- Because of orthogonality, this also applies to the squared Euclidean norm (according to the Pythagorean theorem)

$$\|\mathbf{u}\|^2 = \|\mathbf{P}\mathbf{u}\|^2 + \|\mathbf{Q}\mathbf{u}\|^2$$

Optimal projections and subspaces

- Orthogonal decomposition of squared distances btw. vectors

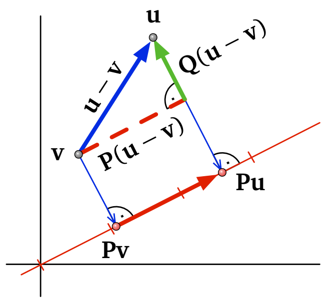
$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{P}\mathbf{u} - \mathbf{P}\mathbf{v}\|^2 + \|\mathbf{Q}\mathbf{u} - \mathbf{Q}\mathbf{v}\|^2$$

- Define projection **loss** as difference btw. squared distances

$$\left| \|\mathbf{P}(\mathbf{u} - \mathbf{v})\|^2 - \|\mathbf{u} - \mathbf{v}\|^2 \right| = \|\mathbf{u} - \mathbf{v}\|^2 - \|\mathbf{P}(\mathbf{u} - \mathbf{v})\|^2 = \|\mathbf{Q}(\mathbf{u} - \mathbf{v})\|^2$$

- Projection quality measure:

$$R^2 = \frac{\|\mathbf{P}(\mathbf{u} - \mathbf{v})\|^2}{\|\mathbf{u} - \mathbf{v}\|^2}$$



References I

Biber, Douglas (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Biber, Douglas (1993). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, **26**, 331–345.

De Sutter, Gert; Delaere, Isabelle; Plevoets, Koen (2012). Lexical lectometry in corpus-based translation studies: combining profile-based correspondence analysis and logistic regression modeling. In M. P. Oakes and J. Meng (eds.), *Quantitative methods in corpus-based translation studies: a practical guide to descriptive translation research*, volume 51 of *Studies in Corpus Linguistics*, pages 325–345. John Benjamins.

Evert, Stefan and Neumann, Stella (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In G. De Sutter, M.-A. Lefer, and I. Delaere (eds.), *Empirical Translation Studies. New Theoretical and Methodological Traditions*, number 300 in Trends in Linguistics. Studies and Monographs (TiLSM), pages 47–80. Mouton de Gruyter, Berlin.

References II

Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Reger, Isabella; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, **22**(suppl_2), ii4–ii16.

Greenacre, Michael (2007). *Correspondence Analysis in Practice*. Interdisciplinary Statistics Series. Chapman & Hall, CRC, 2nd edition.

Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2015). Improving Burrows' Delta. An empirical evaluation of text distance measures. In *Proceedings of the Digital Humanities Conference 2015*, Sydney, Australia.

Jenset, Gard B. and McGillivray, Barbara (2012). Multivariate analyses of affix productivity in translated english. In M. P. Oakes and J. Meng (eds.), *Quantitative methods in corpus-based translation studies: a practical guide to descriptive translation research*, volume 51 of *Studies in Corpus Linguistics*, pages 301–324. John Benjamins.

Perek, Florent (2018). Recent change in the productivity and schematicity of the *way*-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, **14**(1), 65–97.

References III

Sagi, Eyal; Kaufmann, Stefan; Clark, Brady (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 104–111, Athens, Greece.

Speelman, Dirk; Grondelaers, Stefan; Geeraerts, Dirk (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities*, **37**, 317–337.

Tummers, José; Speelman, Dirk; Geeraerts, Dirk (2014). Spurious effects in variational corpus linguistics: Identification and implications of confounding. *International Journal of Corpus Linguistics*, **19**(4), 478–504.