useR! 2014, UCLA

Permutation Tests in MDS

Patrick Mair
Harvard University

Jan De Leeuw
UCLA

Ingwer Borg
GESIS

MDS in a Nutshell

Exploratory technique that maps proximity data of objects into distances between points of a multidimensional space with a given dimensionality p.

- Dissimilarity matrix Δ of dimension $n \times n$ with elements δ_{ij}.
- Problem to solve: Locate points (*configurations*) X in a p-dimensional space such that the distances $d_{ij}(X)$ between the points approximate δ_{ij}.
- Configuration distances:

$$d_{ij}(X) = \sqrt{\sum_{s=1}^{p} (x_{is} - x_{js})^2}$$

- Minimize *stress* (SMACOF uses *Majorization*):

$$\sigma(X) = \sum_{i<j} w_{ij}(\delta_{ij} - d_{ij}(X))^2 \rightarrow \min!$$

MDS in R: smacof Package

`smacof` (De Leeuw & Mair, 2009) allows to fit a variety of MDS models and variants (v.1.5-0):

- simple MDS, spherical MDS, constrained MDS (with optimal scaling on external constraints), individual difference scaling, unfolding.

MDS in R: smacof Package

`smacof` (De Leeuw & Mair, 2009) allows to fit a variety of MDS models and variants (v.1.5-0):

- simple MDS, spherical MDS, constrained MDS (with optimal scaling on external constraints), individual difference scaling, unfolding.
- ratio, interval, ordinal dissimilarities.

MDS in R: smacof Package

`smacof` (De Leeuw & Mair, 2009) allows to fit a variety of MDS models and variants (v.1.5-0):

- simple MDS, spherical MDS, constrained MDS (with optimal scaling on external constraints), individual difference scaling, unfolding.
- ratio, interval, ordinal dissimilarities.
- jackknife and permutation approaches.

In this talk we focus on permutation approaches.

Permutation Approaches to MDS

Make significance statement with respect to a "null configuration". What is a good null configuration?

- Random dissimilarities, nonmetric MDS (Stenson & Knoll, 1969; Spence & Ogilvie, 1973): "nullest of all null hypotheses".
- De Leeuw & Stoop (1984) upper stress bounds, concentric ("degenerate") solution.

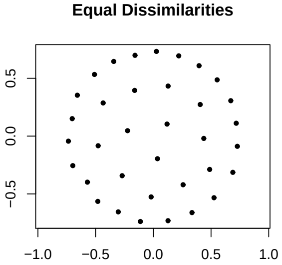
Permutation Approaches to MDS

Make significance statement with respect to a “null configuration”. What is a good null configuration?

- Random dissimilarities, nonmetric MDS (Stenson & Knoll, 1969; Spence & Ogilvie, 1973): “nullest of all null hypotheses”.
- De Leeuw & Stoop (1984) upper stress bounds, concentric (“degenerate”) solution.

Degenerate solution:

- solution with largest stress value.
- stress remains constant across dissimilarity permutations.
- “worst case” solution in terms of structuredness.



Permutation Approaches to MDS

We could now think of the following simulation/permutation strategies:

- Random dissimilarities:
  - Where to draw the dissimilarities from?
  - Test not sharp; basically always significant result.

Permutation Approaches to MDS

We could now think of the following simulation/permutation strategies:

- Random dissimilarities:
  - Where to draw the dissimilarities from?
  - Test not sharp; basically always significant result.
- Permuting Δ:
  - Works well for metric MDS.
  - For nonmetric MDS we end up with the Spence & Ogilvie (1973) standards.
  - Implemented in the `permtest()` function.

Permutation Approaches to MDS

We could now think of the following simulation/permutation strategies:

- Random dissimilarities:
  - Where to draw the dissimilarities from?
  - Test not sharp; basically always significant result.
- Permuting Δ:
  - Works well for metric MDS.
  - For nonmetric MDS we end up with the Spence & Ogilvie (1973) standards.
  - Implemented in the `permtest()` function.
- Permuting original data:
  - Works well if the dissimilarities are computed on the base of a subject $\times$ variable data frame.
  - Row-wise permutation if we want to scale variables.

Permutation Approaches to MDS

We could now think of the following simulation/permutation strategies:

- Random dissimilarities:
  - Where to draw the dissimilarities from?
  - Test not sharp; basically always significant result.
- Permuting Δ:
  - Works well for metric MDS.
  - For nonmetric MDS we end up with the Spence & Ogilvie (1973) standards.
  - Implemented in the `permtest()` function.
- Permuting original data:
  - Works well if the dissimilarities are computed on the base of a subject $\times$ variable data frame.
  - Row-wise permutation if we want to scale variables.
- Mantel-type test (Mantel, 1967; Legendre & Fortin, 1989):
  - Permutation test on whether 2 dissimilarity matrices are equal.
  - One matrix is Δ, the other one contains constant dissimilarities.

Example: Republican Statements

We've scraped statements from the GOP website (`www.gop.com`) where voters had to complete the sentence "I am a Repbublican because ...".

"... I stand for freedom, limited government, fiscal responsibility, and keeping the USA the Greatest Country on Earth."
"... I believe that America represents the greatest ideals and hopes of mankind."
"... I believe in small government, big military, and in the traditional core family values."
"... I believe in low taxes, strong national defense. right to bear arms, right to life, and no government run health care."
"... I believe in a free market society which enables hard work to equal success – I am also very pro life and against same sex unions."

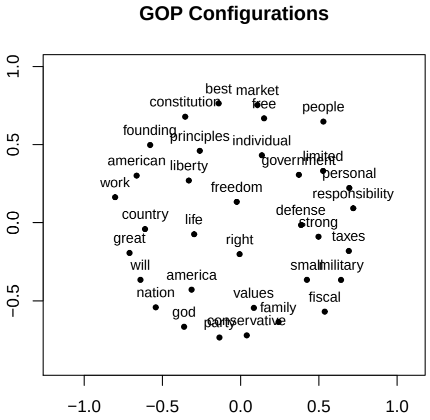
Questions: How are the terms voters use associated with each other? Can we find word clusters that represent value structures related to certain Republican subgroups?

Analysis:

- DTM of the 35 most frequent words across 254 statements.
- Cosine distances between word frequency vectors $\rightarrow \Delta$.
- Metric MDS on Δ using `smacofSym()` (2D solution).

GOP: MDS solution

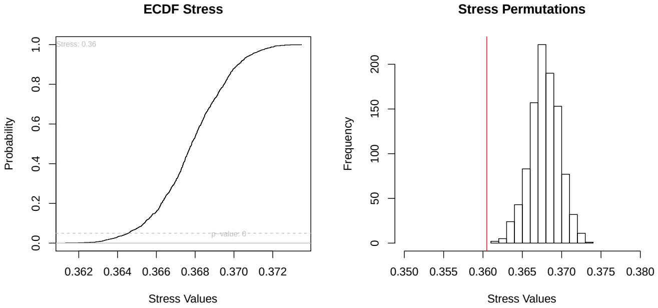
We get the following configuration plot (stress = 0.3605):



GOP Fit I: Permute Dissimilarities Δ

First let's permute Δ (1000 times) and compute a SMACOF solution for each Δ_i.

H_0: dissimilarities random.

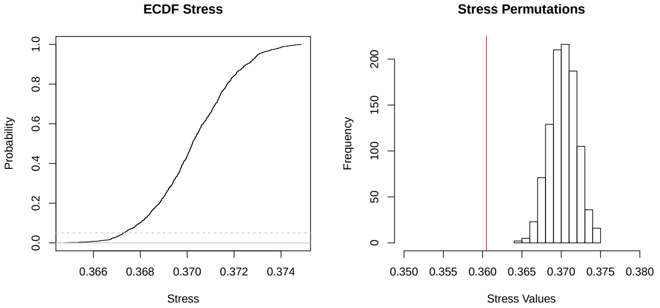


We get a p-value of 0.000.

GOP Fit II: Permute Data (DTM)

Let's now perform row-wise permutations of the DTM, compute cosine distances (column-wise), fit SMACOF on each Δ_i.

H_0: no differences across variables.



We get a p-value of 0.000.

GOP Fit III: Mantel-type Test

Mantel test: permute between observed Δ and Δ_0 (constant dissimilarities).

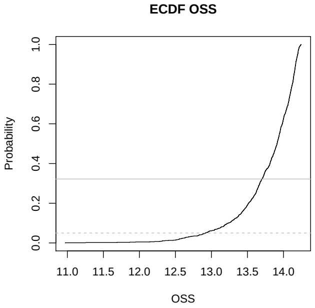
- SMACOF on $\Delta_i^{(1)}$ and $\Delta_i^{(2)}$.
- Procrustes on configurations ($X_i^{(1)}$, $X_i^{(2)}$).
- Squared differences between Procrustes matrices as test statistic.

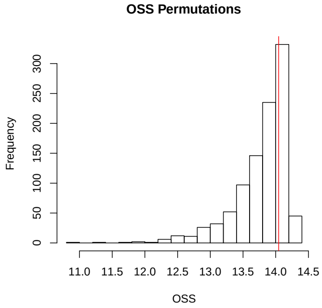
GOP Fit III: Mantel-type Test

Mantel test: permute between observed Δ and Δ_0 (constant dissimilarities).

- SMACOF on $\Delta_i^{(1)}$ and $\Delta_i^{(2)}$.
- Procrustes on configurations $(X_i^{(1)}, X_i^{(2)})$.
- Squared differences between Procrustes matrices as test statistic.

H_0: no difference between Δ and Δ_0.





We get a p-value of 0.322.

Summary and Outlook

Three types of permutation tests:

- Both, permuting the dissimilarities and permuting the original data test randomness hypotheses.
- Mantel-type permutation tests are tests on structuredness:
  - Various structural hypotheses can be tested.
  - Various test statistics for comparing the two matrices can be considered.
  - Performance needs to be studies in more detail, however.

All three types are can be applied to constrained MDS variants and individual difference scaling as well.

For unfolding models: within rows permutations on input preference matrix.

References

Package:

De Leeuw, J. & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software, 31(3)*, p. 1–30.

Permutation:

Stenson, H. H. & Knoll, R. L. (1969). Goodness of fit for random rankings in Kruskal's nonmetric scaling procedures. *Psychological Bulletin, 72*, 122–126.

Spence, I., & Ogilvie, J. C. (1973). A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivariate Behavioral Research, 8*, 511–517.

De Leeuw, J. & Stoop, I. (1984). Upper bounds for Kruskal's stress. *Psychometrika, 49*, 391–402.

Mantel test:

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209–220.

Legendre, P., & Fortin, M. (1989). Spatial pattern and ecological analysis. *Vegetatio, 80*, 107–138