Thomas Rusch and Patrick Mair and Kurt Hornik

The STOPS framework for structure-based hyperparameter selection in multidimensional scaling

Conference or Workshop Item (Published)
(Refereed)

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

The STOPS Framework

for Structure-Based Hyperparameter Selection in Multidimensional Scaling

Slide Zero

This is joint work with Patrick Mair (Harvard) and Kurt Hornik (WU)

Multidimensional Scaling

The STRESS objective function with (transformed) distances $d^*_{ij}(X)$, (transformed) proximities δ^*_{ij} and finite weights w^*_{ij} is

$$\sigma(X) = \sum_{i<j} w^*_{ij} \left[\delta^*_{ij} - d^*_{ij}(X)\right]^2$$

which is minimized to find the configuration X

$$\arg\min_X \sigma(X)$$

- MDS provides an optimal map into continuous space $\mathbb{R}^M$ (objective 1)
- We may also be interested in some structural appearance of X, e.g., clusters or circumplex (objective 2).
- It can happen that what is optimal for objective 1 is not very useful for objective 2

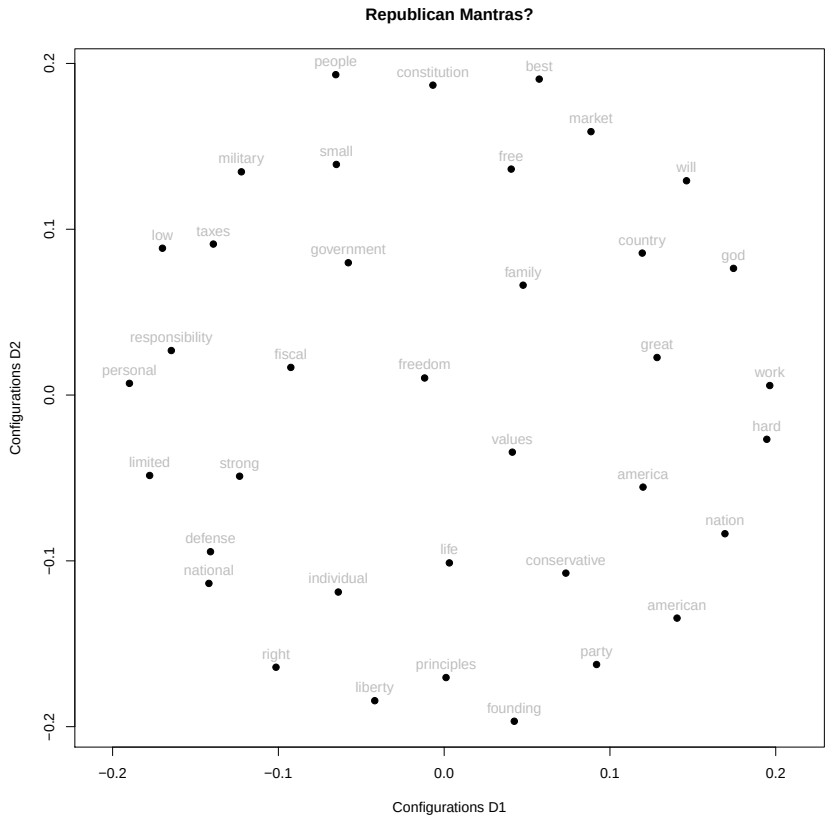
Motivation: Republican Mantras

"I'm a Republican, because ..." from Mair et al. (2014)

- Supporters of the Republican Party have been asked why they are Republican (254 statements)
- Natural language data that was scraped and processed $\Longrightarrow$ Sparse data matrix (document term matrix)
- Objects are the words (we use only words that appeared at least 10 times)
- We look for themes in the statements: "Mantras" (words that occur often together)

We use a cosine distance for word co-occurences and apply standard least squares MDS (SMACOF) for representation.

Motivation: Republican Mantras



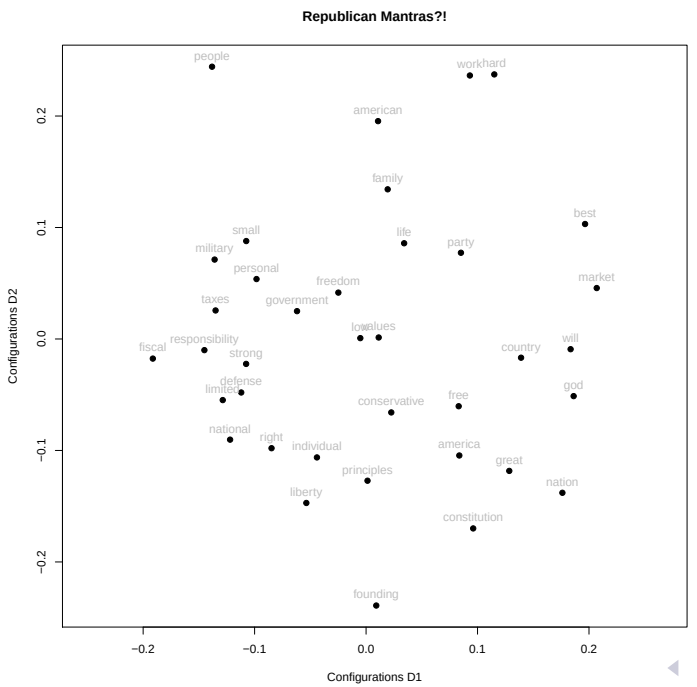
- We find lack of (interesting) structure in MDS configuration

Multidimensional Scaling Extensions

- More structure is often introduced by using transformations $\delta_{ij}^* = f_{ij}(\delta_{ij})$ and $d_{ij}(X)^* = g_{ij}(d_{ij}(X))$ and weights w_{ij}^*
- Many MDS variants are a special case of this general formulation, e.g.,
  - Metric MDS: $g_{ij}(a) = a$, $f_{ij}(a) = a$, Sammon mapping: $w_{ij}^* = \delta_{ij}^{-1}$
  - Multiscale: $f_{ij}(a) = g_{ij}(a) = \log(a)$
  - POST-MDS: $g_{ij}(a) = a^\kappa, f_{ij}(a) = a^\lambda, w_{ij}^* = w_{ij}^\nu$, ALSCAL: $\kappa = \lambda = 2$
  - LMDS: Box-Cox transformations for $g_{ij}(\cdot), f_{ij}(\cdot)$, Isomap: $g_{ij}(\cdot)$ isometric distance
- Often transformations are parametrized by a hyperparameter vector θ, so $\delta_{ij}^* = f_{ij}(\delta_{ij}; \theta)$ and $d_{ij}^* = g_{ij}(d_{ij}; \theta)$

Power Stress MDS

- Fit ratio MDS with power transformation by setting, e.g., $f(\delta_{ij}) = \delta_{ij}^{20}$
- Structure is clearer but the fit is now worse (0.373 versus 0.401) (essentially fits only δ very close to the maximum)
- Parameters chosen *ad hoc*, not always clear what is the right θ.



Structure Optimized Proximity Scaling

Our suggestion is a framework to systemize this approach: Structure Optimized Proximity Scaling (STOPS).

- Idea: Select the parameters for the transformations (θ) in a principled fashion by fit and structure considerations
- This offers a conceptual and computational framework for hyperparameter selection in MDS variants
- Building blocks:
  - θ–parametrized target function for misfit
  - Statistics measuring configuration structure (structuredness indices)
  - Combination of misfit and structure
  - Algorithm for optimization

STOPS - I

We have the target function that measures misfit (e.g., Stress)

$$\sigma(X, \theta) = L(\Delta^*, D^*(X), \theta)$$

which we minimize to find the configuration X for a θ

$$X(\theta) = \arg\min_X \sigma(X, \theta)$$

- $X(\theta)$ has some structural appearance (C-Structuredness).
- C-Structuredness changes with different θ

STOPS - II

- Capture P structures in $X(\theta)$ by indices $I_p(X(\theta);\gamma), p = 1, \ldots, P$.
- Combine $\sigma(X(\theta),\theta)$ and $I_p(X(\theta);\gamma)$ to stoploss$(X(\theta),\vartheta;\Delta)$
- Two STOPS models
  - Additive STOPS (aSTOPS)

$$\text{stoploss}(X(\theta),\vartheta;\Delta) = v_0 \cdot \sigma(X(\theta),\theta) + \sum_{p=1}^{P} v_p I_p(X(\theta);\gamma)$$

  - Multiplicative STOPS (mSTOPS)

$$\text{stoploss}(X(\theta),\vartheta;\Delta) = \sigma(X(\theta),\theta)^{v_0} \cdot \prod_{p=1}^{P} I_p(X(\theta);\gamma)^{v_p}$$

v_0.. stressweight (redundant), $v_1, ..., v_P$... structuredness weights, γ... (optional) metaparameters for structuredness indices; $\vartheta \subseteq \{\theta, v_0, \ldots, v_k\}$

Structures and Indices

- C-Structuredness indices capture essence of a particular structure in a configuration. Some examples:
  - C-Association: Pairwise nonlinear association between principal axes (pairwise maximal maximum information coefficient; Reshef et al. 2011)
  - C-Clusteredness: A clustered appearance (normed OPTICS Cordillera; Rusch et al., 2018)
  - C-Complexity: Complexity of the functional relationship between any principal axes (pairwise maximal minimum cell number; Reshef et al. 2011)
  - C-Manifoldness: Points lie close to a smooth submanifold (maximal correlation; Sarmanov, 1958)

Optimization-I

We need to find

$$\arg\min_{\vartheta} \ \text{stoploss}(X(\theta), \vartheta; \Delta)$$

- This can be seen as a profile method
- We use a nested algorithm
  1. First solve for $X(\theta) = \arg\max_X \sigma(X, \theta)$
  2. Then minimize $\text{stoploss}(X(\theta), \vartheta; \Delta)$ over ϑ
- Advantages:
  - For finding $X(\theta)$ we can use standard solutions (reasonably good)
  - The inner part (1.) allows computationally flexible specifications of MDS method
  - $I_p(X)$ depends directly only on $X(\theta)$
  - Dimensionality of outer problem is usually not very high

Optimization-II

- Difficulties when optimizing over ϑ
  - Inner minimization is very costly
  - For stoploss basically only know function evaluations
  - Estimation of Step 1 may be noisy (premature termination, local minimum)
- This suggests to solve Step 2 with Efficient Global Optimization aka Bayesian Optimization.
- One samples the “best” candidate for evaluation given a surrogate model and the current knowledge.

Optimization-III

- Bayesian Optimization:
  - Choose a (flexible) surrogate model (prior)
  - Evaluate the target function at some candidate values (data)
  - Update the prior with the function evaluations (posterior)
  - Maximize an acquisition function over the posterior surface
  - This suggests a candidate parameter combination
  - Evaluate at candidate and repeat
- We use Expected Improvement for acquisition and Treed Gaussian Process with Jumps to Linear Models (Grammacy, 2007) or Kriging (Roustant et al., 2012) for the surrogate model.

R Package stops

All of this is implemented in the R package stops

- High level function for STOPS `stops(delta,loss,...)`
- Prespecified MDS models (argument `loss`) are `strain`, SMACOF (`smacofSym`), `sammon` mapping, `elastic` scaling, SMACOF on a sphere (`smacofSphere`), `sstress`, `rstress`, `powerstress`, Sammon mapping and elastic scaling with powers (`powersammon`, `powerelastic`). Planned: Isomap and LMDS
- Optimization with Bayesian optimization (`kriging`, `tgp`) and some more (including simulated annealing `SANN` or a particle swarm algorithm `pso`).
- Features various c-structuredness indices
- S3 methods: `plot, summary, print, coef, residuals, plot3d, plot3dstatic`

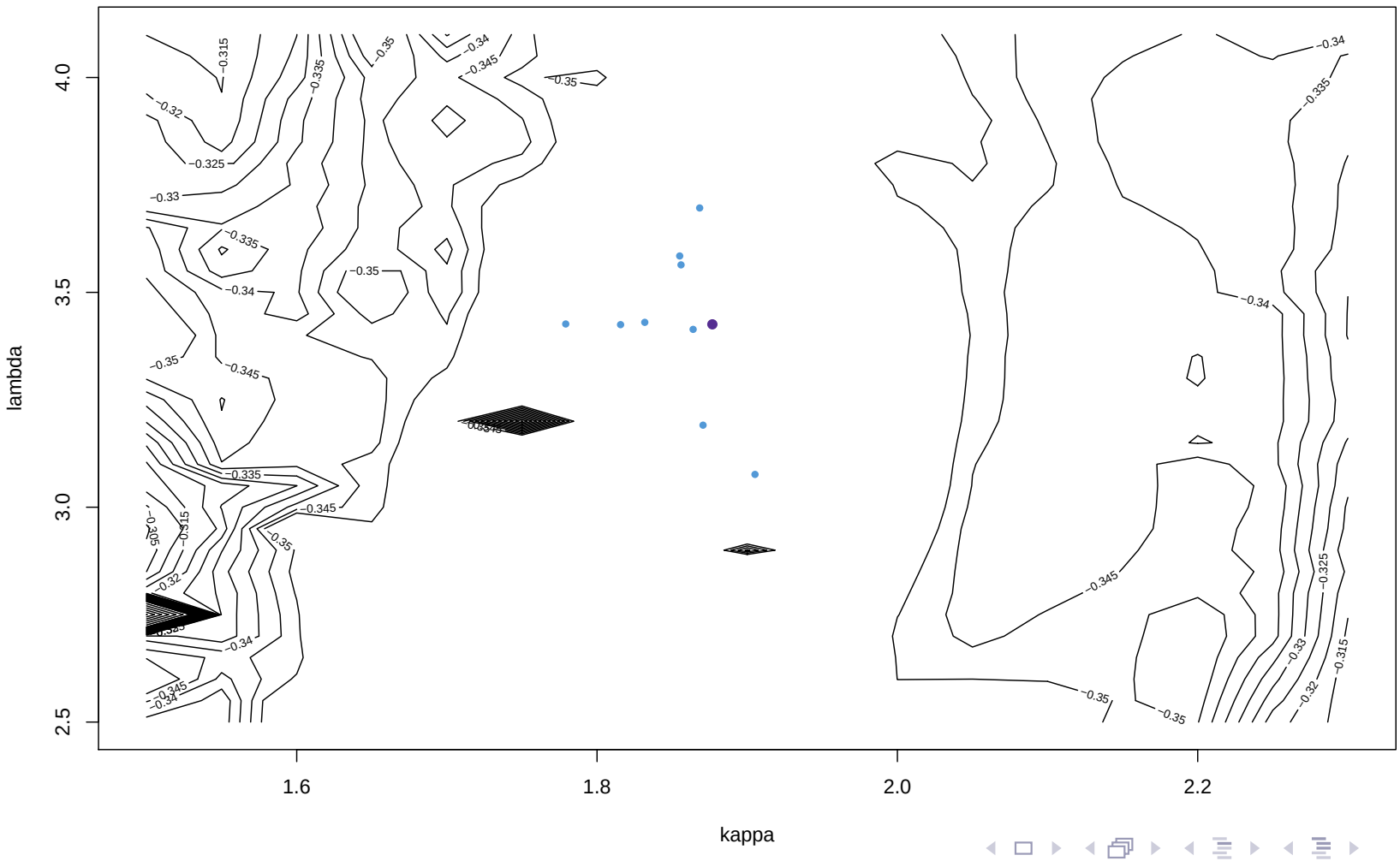
Example: Republicans

- Misfit: Power Stress MDS
- Structuredness: C-Clusteredness and C-Manifoldness
- Optimization with treed gaussian process prior with jump to linear models (for 20 steps)

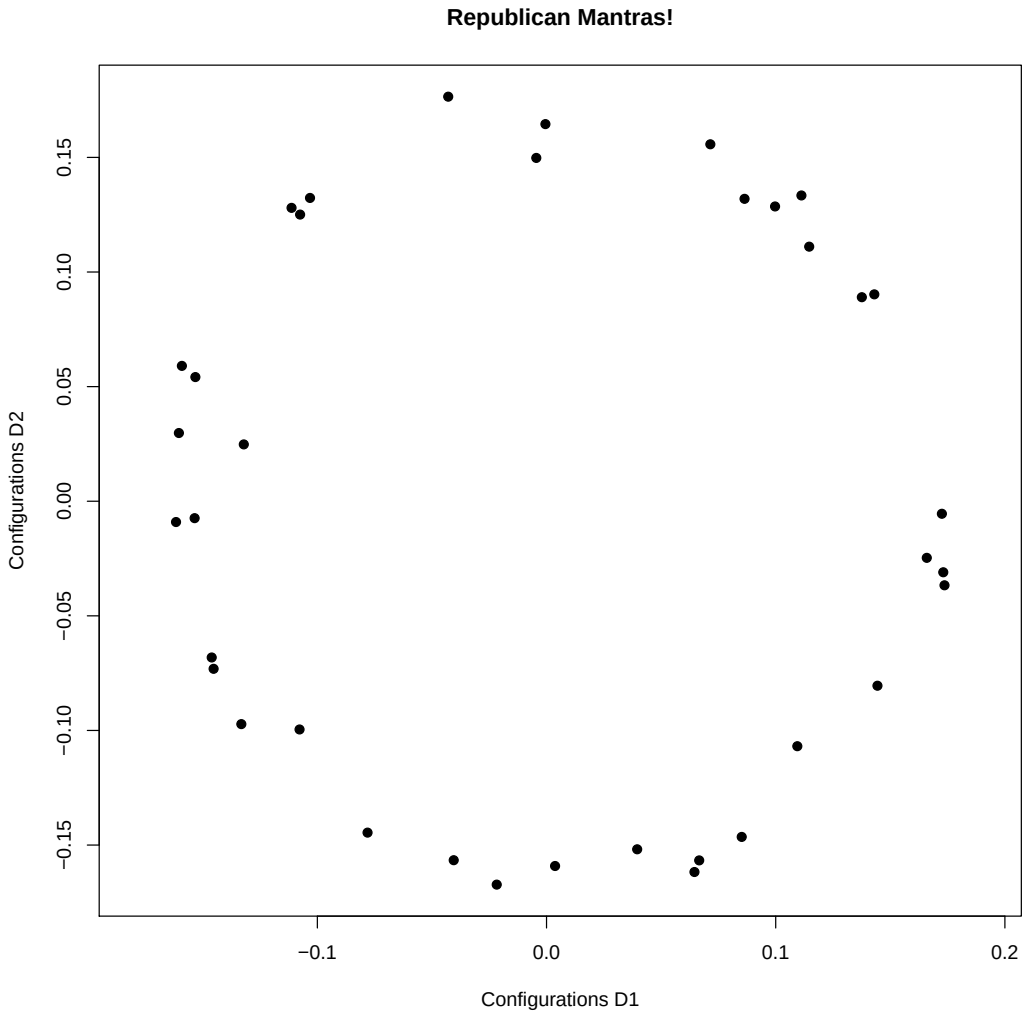
```
R> resc <- stops(dt.dist,loss="powermds",
+                structures=c("cmanifoldness","cclusteredness"))
R> resc
```

```
Call: stops(dis = dt.dist, loss = "powermds", theta = c(1, 1), structures = c("cmanifoldness",
    "cclusteredness"), strucpars = strucpars, optimmethod = "tgp",
    lower = c(0.5, 0.3), upper = c(3, 10), verbose = 5, type = "additive",
    itmax = 20)

Model: additive STOPS with powermds loss function and theta parameters= 1.871 3.191 1

Number of objects: 37
MDS loss value: 0.2513
C-Structuredness Indices: cmanifoldness 0.9738 cclusteredness 0.3117
Structure optimized loss (stoploss): -0.3914
MDS loss weight: 1 c-structuredness weights: -0.5 -0.5
Number of iterations of tgp optimization: 20
```

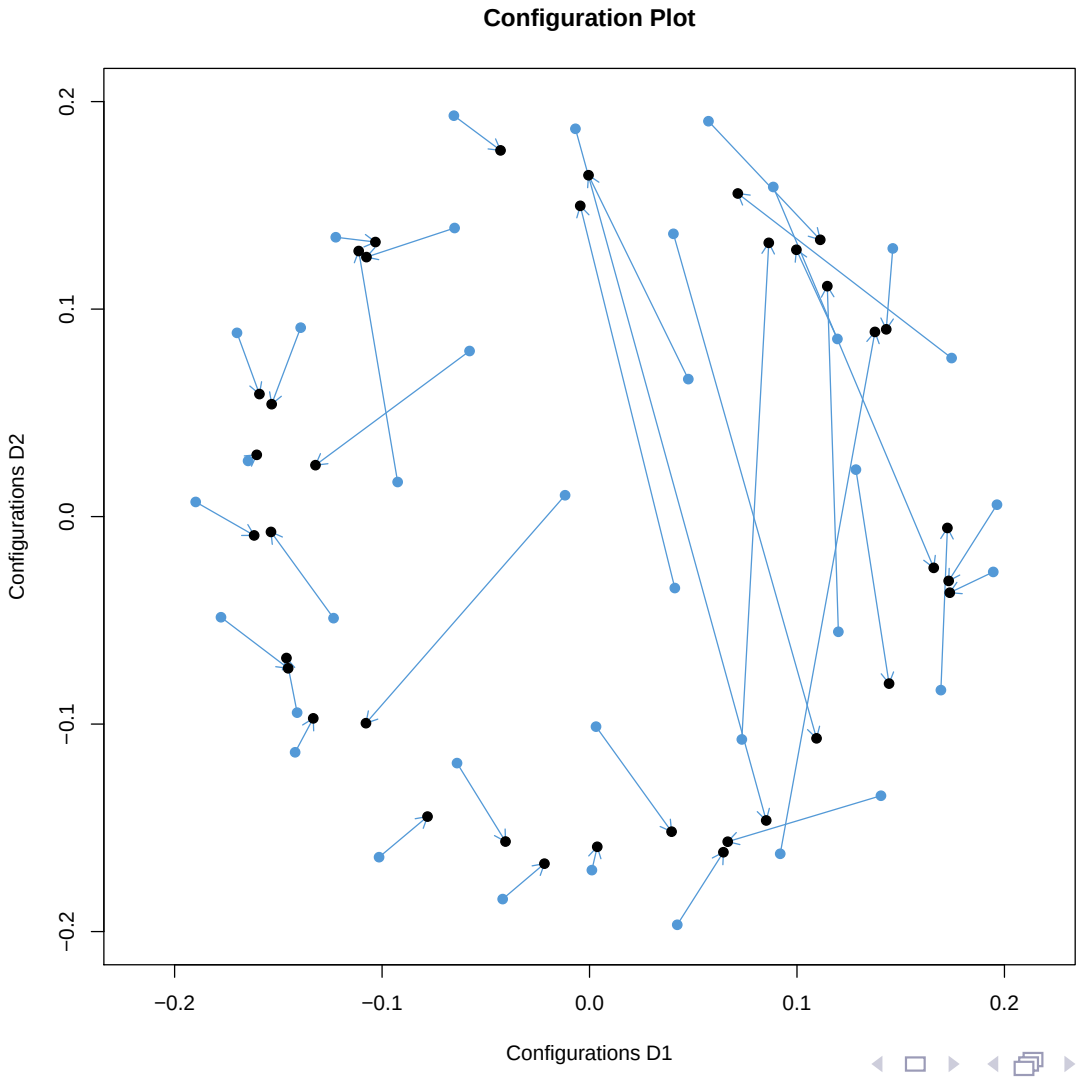
Example: Republicans



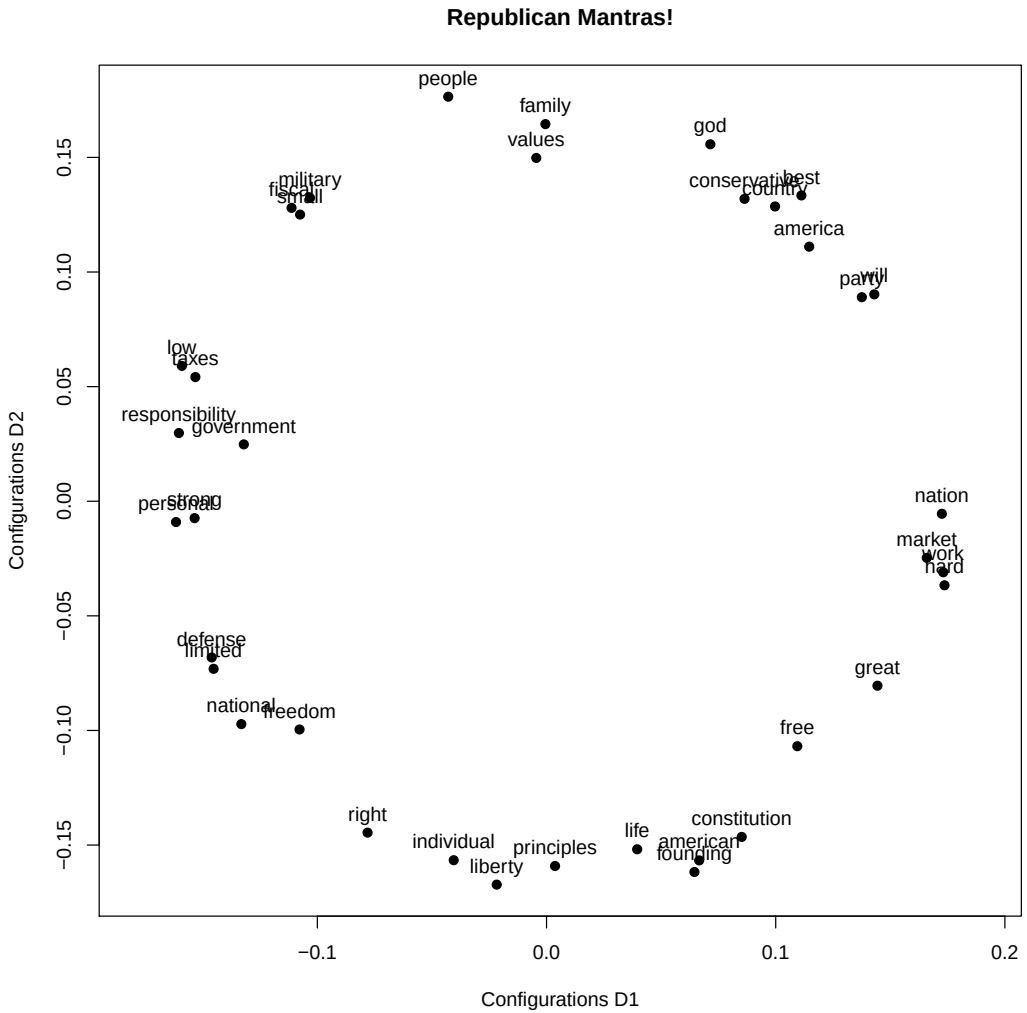
Example: Republicans



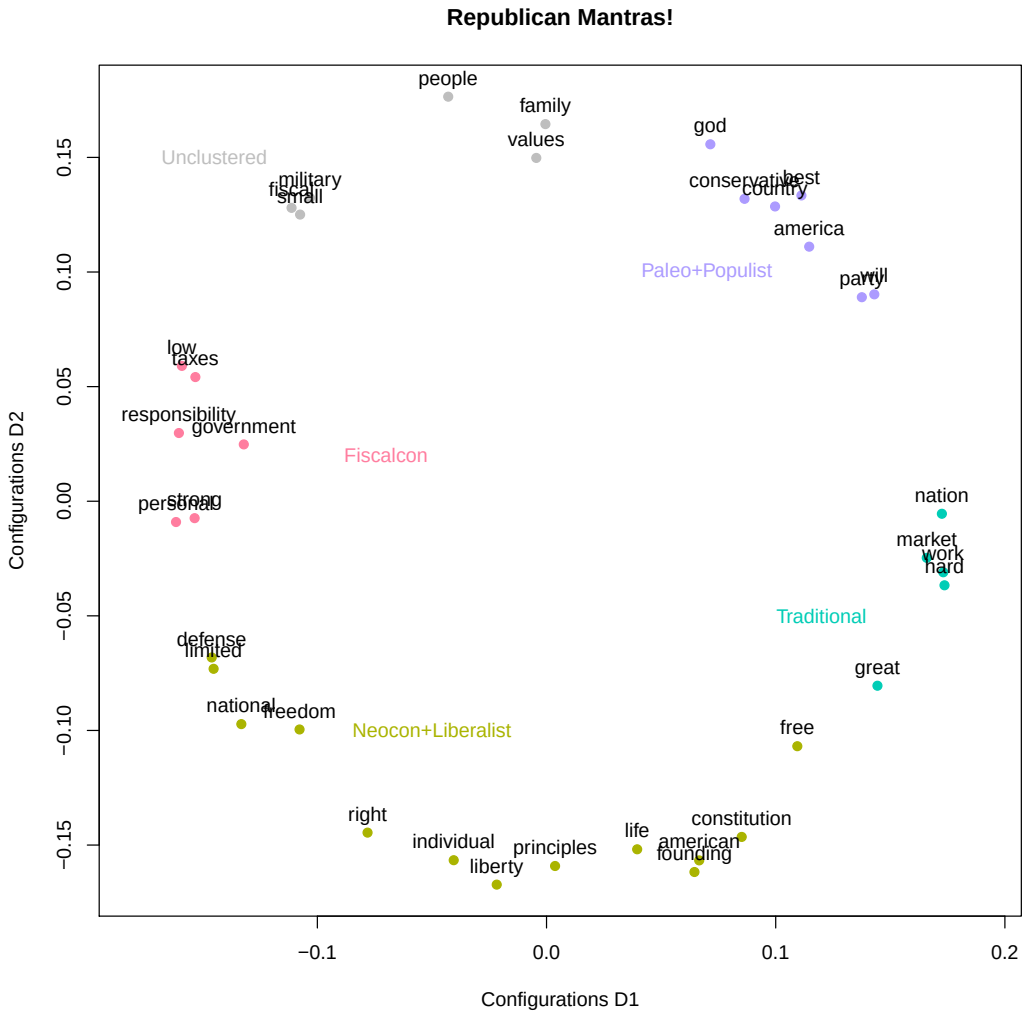
Example: Republicans



Example: Republicans



Example: Republicans



Summary and Outlook

STOPS

- A conceptual and computational framework for hyperparameter optimization in MDS based on structure considerations

Outlook

- More models and (perhaps?) more structures
- Extend to other dimension reduction techniques (e.g., the Gifi system)

References

- Borg, I., Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*, 2nd Edition, Springer, New York.
- Gramacy, R. B. (2007). tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *Journal of Statistical Software, 19(9)*, 1–46.
- Mair, P., Rusch, T., Hornik, K. (2014) The grand old party - A party of values? *SpringerPlus*, 3:697.
- Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., & Sabeti, P. (2011) Detecting novel associations in large data sets. *Science, 334*, 1518–1524.
- Roustant, O., Ginsbourger, D., & Deville, Y. (2012). Dicekriging, Diceoptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization. *Journal of Statistical Software, 51(1)*, 1–54.
- Rusch, T., Hornik, K., Mair, P. (2018) Assessing and quantifying clusteredness: The OPTICS Cordillera. *Journal of Computational and Graphical Statistics, 27 (1)*, 220-233.
- Rusch, T., Mair, P., Hornik, K. (in preparation). Structure based hyperparameter selection for dimensionality reduction: The STOPS framework for Structure Optimized Proximity Scaling.
- Sarmanov, O. (1958). Maximum correlation coefficient (symmetric case). Doklady Akad. Nauk SSR, 120, 715–718.

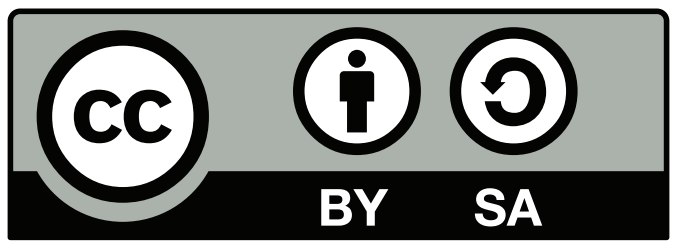
Thank You for Your Attention

Thomas Rusch
Competence Center for Empirical Research Methods
email: thomas.rusch@wu.ac.at
URL: http://wu.ac.at/methods/team/dr-thomas-rusch

WU Vienna University of Economics and Business
Welthandelsplatz 1, 1020 Vienna
Austria

License

Please attribute Thomas Rusch, Patrick Mair and Kurt Hornik. Except where otherwise noted, this work is licensed under CC-BY-SA:



https://creativecommons.org/licenses/by-sa/4.0/