WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Structure Optimized Proximity Scaling (STOPS)

A Framework for Hyperparameter Selection in Multidimensional Scaling

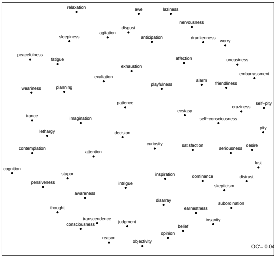
Slide Zero

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

Motivation: Mental States

- Tamir et al. (2016) investigated how our brain represents the mind of others (social cognition) by correlation of activation patterns of fMRI brain scans
  - For 20 individuals and 60 mental states
  - Task was to choose for a given mental state the one out of two situations most likely to induce the state in others
  - In supplement the authors invite readers to explore the neural similarity of states directly by means of Multidimensional Scaling (MDS)
- We average correlation-derived dissimilarities over the 20 individuals.

Motivation: Lack of Structure



- We find lack of structure in MDS (Spherical Embedded Projection Phenomenon)
- Characterized by objects arranged on a disk or sphere and not uncommon
- Appears when the observed proximities δ_{ij} have little variability

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, e.g., clusters or circumplex (objective 2).
- It can happen that what is optimal for objective 1 is not very useful for objective 2

Multidimensional Scaling Extensions

- Structure often becomes clearer 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 $d_{ij}^* = g_{ij}(d_{ij}; \theta)$
- It is not always clear what is the right θ.

Structure Optimized Proximity Scaling

Our suggestion is 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 $I_p(X(\theta);\gamma), p = 1, \dots, P$.
- 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), $v_1, \dots, v_P$... structuredness weights, γ... (optional) metaparameters for structuredness indices; $\vartheta \subseteq \{\theta, v_0, \dots, v_k\}$

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., 2016)
  - C-Complexity: Complexity of the functional relationship between any principle axes (pairwise maximal minimum cell number; Reshef et al. 2011)
  - C-Dependence: Random vectors of projections onto the axes are stochastically dependent (distance correlation; Szekely et al., 2007)
  - 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`

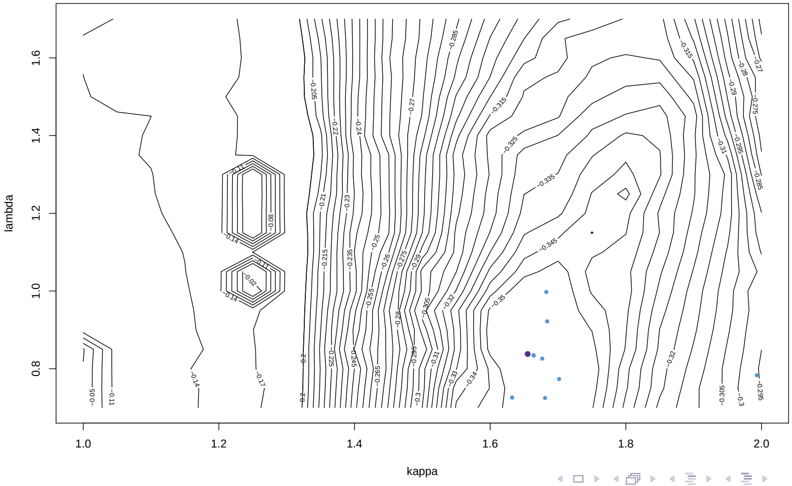
Example: Mental States - I

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

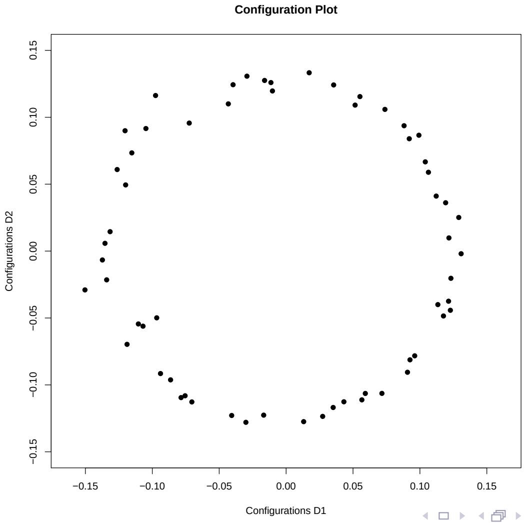
```
R> res1 <- stops(dis,loss="powermds",theta=c(1,1,1),structures=c("ccluste
+         "cmanifoldness"),optimmethod="tgp",itmax=20,lower=c(1,0.7,1),upp
R> res1
```

```
Call: stops(dis = dis, loss = "powermds", theta = c(1, 1, 1), structures = c("cclusteredness",
    "cmanifoldness"), optimmethod = "tgp", lower = c(1, 0.7,
    1), upper = c(2, 5, 1.1), verbose = 5, initpoints = 10, itmax = 20)

Model: additive STOPS with powermds loss function and theta parameters= 1.677 0.826 1

Number of objects: 60
MDS loss value: 0.2539
C-Structuredness Indices: cclusteredness 0.2588 cmanifoldness 0.9664
Structure optimized loss (stoploss): -0.3587
MDS loss weight: 1 c-structuredness weights: -0.5 -0.5
Number of iterations of tgp optimization: 20
```

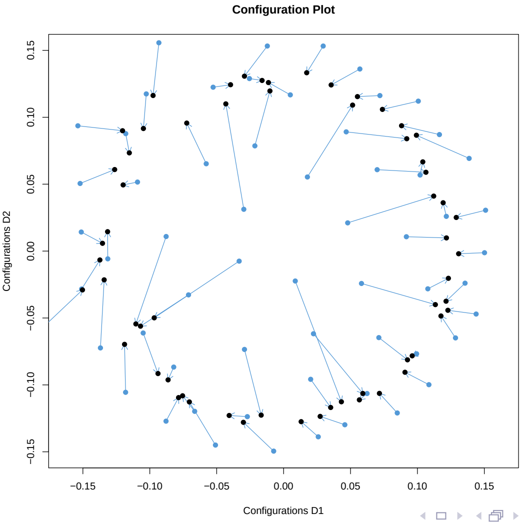
Example: Mental States - II



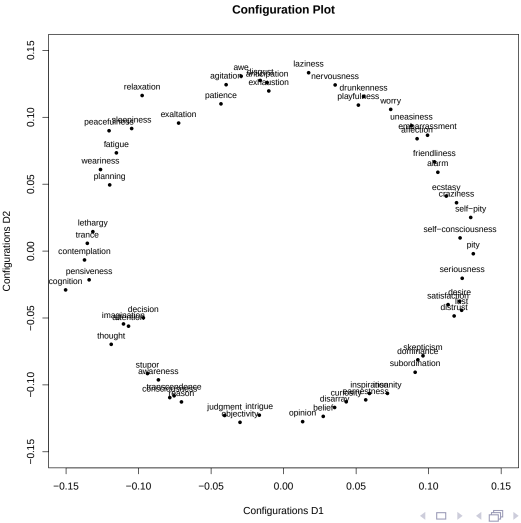
Example: Mental States - III



Example: Mental States - IV



Example: Mental States - V



Summary and Outlook

STOPS

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

Outlook

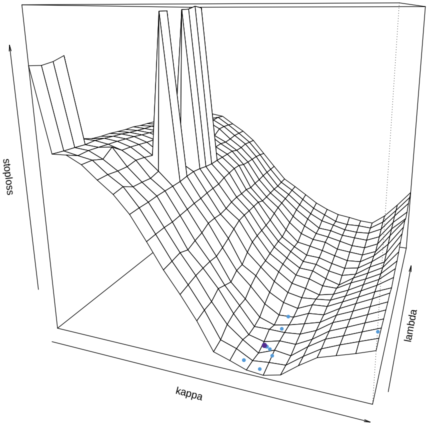
- More models and (perhaps?) more structures
- Extend to other dimension reduction techniques (e.g., the Gifi system)

References

- 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.
- 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. (2016) Assessing and quantifying clusteredness: The OPTICS Cordillera. Report 2016/1, Discussion Paper Series / Center for Empirical Research Methods, 2016/1. WU Vienna University of Economics and Business, Vienna.
- 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.
- Szekely, G., Rizzo, M., & Bakirov, N. (2007). Measuring and Testing Independence by Correlation of Distances, Annals of Statistics, 35 (6), 2769âĂŞ2794.
- Tamir, D.I., Thornton, M.A., Contreras, J.M., & Mitchell, J. P. (2016) Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. PNAS, 113 (1), 194-199.

Backup Slides

Example: Mental States - 3D



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