COPS and STOPS

Cluster and/or Structure Optimized Proximity Scaling

Outline

1 Problem Motivation

2 COPS: Cluster Optimized Proximity Scaling
- C-Clusteredness
- The COPS Procedure
- COPS Variants
- COPS Example

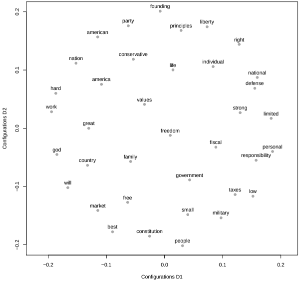
3 STOPS: Structure Optimized Proximity Scaling
- STOPS Framework
- Structuredness Indices
- Optimization
- Package
- STOPS Example

4 Conclusion and Outlook

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

Lack of Structure in MDS

- In exploratory data analysis we may look for anything, but find little to nothing
- E.g., “I’m a Republican, because...” statements (Mair et al., 2014) with MDS on cosine distance between words from co-occurrences.

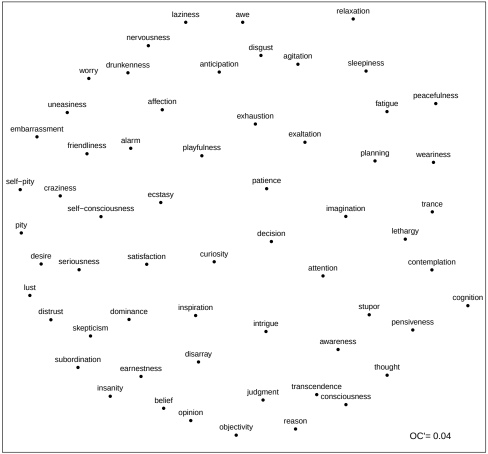


- Looked for word clusters but have a lack of structure

Another Example

- In MDS this is not an uncommon situation (embedded sparse sphere phenomenon)
- Mental States Data: Tamir et al. (2016) investigates 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 2-dim MDS

Neural States MDS



Dimensionality Reduction

- Methods that provide a mapping from a higher dimensional to a lower dimensional space based on some idea of optimality
- Example: Multidimensional Scaling (MDS)
  - Least Squares MDS utilizes the STRESS loss function

$$\sigma_{MDS}(X) = \sum_{i<j} w_{ij}^* \left[f_{ij}(\delta_{ij}) - g_{ij}(d_{ij}(X)) \right]^2$$

  - and minimizes it to find the configuration X

$$\arg\min_X \sigma_{MDS}(X)$$

$d_{ij}(X)$ ... fitted distances, δ_{ij} ... proximities
$g_{ij}(\cdot), f_{ij}(\cdot)$ ... transformation functions
w_{ij}^* ... finite weights

Multidimensional Scaling (MDS)

- Provides an optimal map into continuous space $\mathbb{R}^M$ and looks for directions of spread in the low dimensional space (objective 1)
- But we may be interested in some structural idea, e.g., discrete structures of similarity between objects (“clusters”; objective 2)
- MDS does solve objective 1 but not objective 2. The latter is often inferred from the former by how it looks
- It can happen that what is optimal for objective 1 is not very useful for objective 2
- One way out: Use transformations so clustering is clearer.
- Often this means that the fit may get worse

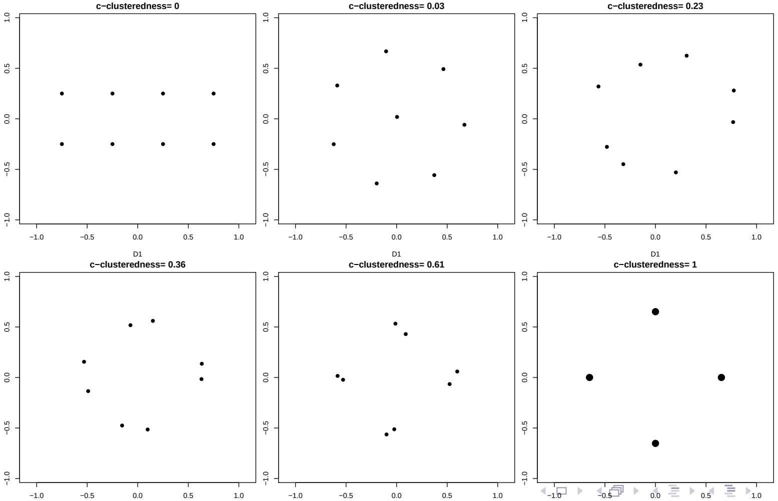
COPS for the Rescue

Our solution to this problem: COPS (*C*luster *O*ptimized *P*roximity *S*caling; Rusch et al., 2015a).

- Use STRESS with θ-parametrized monotonic nonlinear transformations of proximities and/or fitted distances. e.g., power transformations (powerStress, $g(d_{ij}(X)) = d_{ij}(X)^{\kappa}$ and $f(\delta_{ij}) = \delta_{ij}^{\lambda}$, $w_{ij}^{*} = w_{ij}^{\nu}$, so $\theta = c(\kappa, \lambda, \nu)$)
- Use an index of the obtained degree of clusteredness in the configuration (c-clusteredness) to quantify how clustered the result is
- Combine this into a single target function and optimize
- Two versions:
  - COPS-C (Optimize combined loss to get X)
  - P-COPS (Profile method to find θ)

C-Clusteredness

C-Clusteredness: The amount of clusteredness of a configuration



OPTICS Cordillera - I

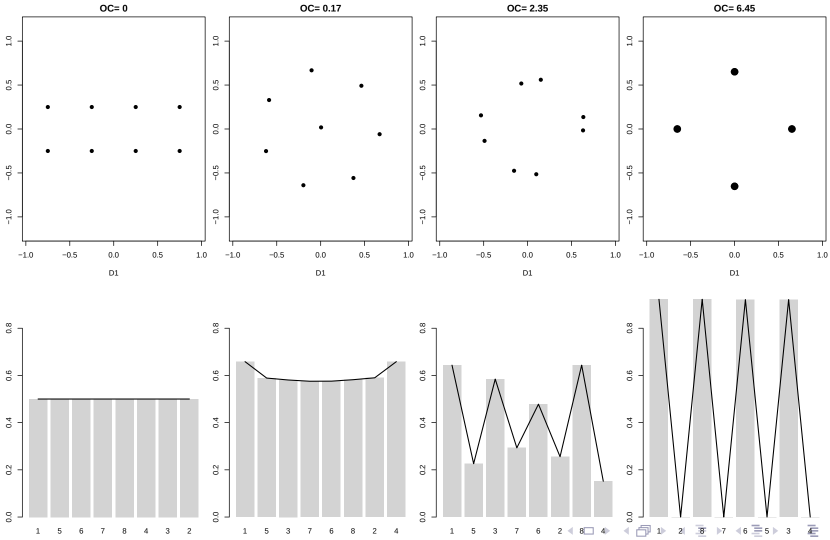
Index for clusteredness: OPTICS Cordillera (Rusch et al., 2016)

- Employs OPTICS (Ankerst et al., 1999) with metaparameters k, ϵ on the configuration distances. For row vectors x_j of X returns an ordering R of these points, $R = \{x_{(i)}\}_{i=1,\dots,N}$.
- OPTICS also returns a reachability plot (dendrogram of minimum reachabilities $r^*_{(i)}$ of point $x_{(i)}$)
- Ordering and reachability represent the clustering structure. We aggregate that to an index $OC'(X)$ by defining (for metaparameter $q > 0$)

$$\mathrm{OC}'(X) = \left(\frac{\sum_{i=2}^{N} \left| r^*_{(i)} - r^*_{(i-1)} \right|^q}{d_{max}^q \cdot \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right)} \right)^{1/q}$$

- It holds that $0 \leq OC'(X) \leq 1$.

OPTICS Cordillera - II



The COPS Procedure

Combine the $\theta-$ parametrized STRESS, $\sigma_{MDS}(X(\theta),\theta)$ and the OPTICS cordillera $OC(X)$ to cluster optimized loss (coploss):

$$\text{coploss}(X,\theta) = v_1 \cdot \sigma_{MDS}(X,\theta) - v_2 \cdot \text{OC}(X) \tag{1}$$

and $v_1, v_2 \in \mathbb{R}$ controlling how much weight should be given to the individual parts of coploss.

We derive two versions from this loss

- COPS-C: $\text{coploss}(X;\theta) = v_1 \cdot \sigma_{MDS}(X;\theta) - v_2 \cdot \text{OC}(X;\theta)$
- P-COPS: $\text{coploss}(\theta) = v_1 \cdot \sigma_{MDS}(X(\theta),\theta) - v_2 \cdot \text{OC}(X(\theta))$ with $X(\theta) := \arg\max_X \sigma(X,\theta)$.

COPS-C

- Using COPS to find a configuration
- We need to do

$$\text{coploss}(X;\theta) \rightarrow min_X!$$

- We use the derivative free heuristic NEWUOA
- Works well when initial configuration is near the optimum
- Set initial configuration X^0 to $min_X \sigma_{MDS}(X)$
- Local improvement towards more c-clusteredness for the MDS solution

P-COPS

- Profile Version of COPS for hyperparameter selection
- We need to do

$$\text{coploss}(\theta) \rightarrow min_\theta!$$

- We use a nested algorithm that first solves for $X(\theta)$ and then minimizes over θ.
  - For the inner part, i.e., finding $X(\theta)$ standard MDS optimization is used (e.g., majorization)
  - The outer part of this optimization problem we use metaheuristics (good experiences with an adapted Luus-Jaakola algorithm (Luus & Jaakola, 1973)

Example: Mental States COPS

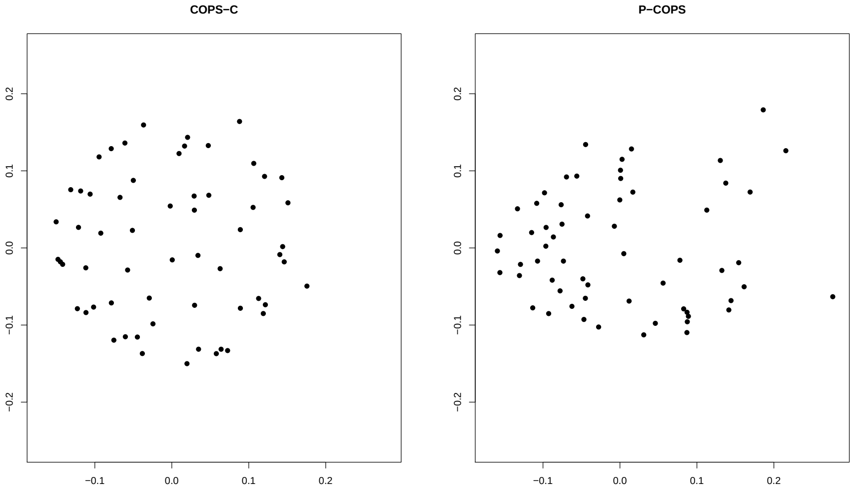
- COPS-C: `cops(dis,'COPS-C',stressweight=0.9,cordweight=0.1)`

```
Call: [1] "[deleted]"

Model: COPS with parameters kappa= 1 lambda= 1 nu= 1

Number of objects: 60
Stress of configuration (default normalization): 0.3671
OPTICS Cordillera: Raw 10.94 Normed 0.2504
Cluster optimized loss (coploss):  0.09625
Stress weight: 0.9  OPTICS Cordillera weight: 0.1
Number of iterations of Newuoa optimization: 13292
```

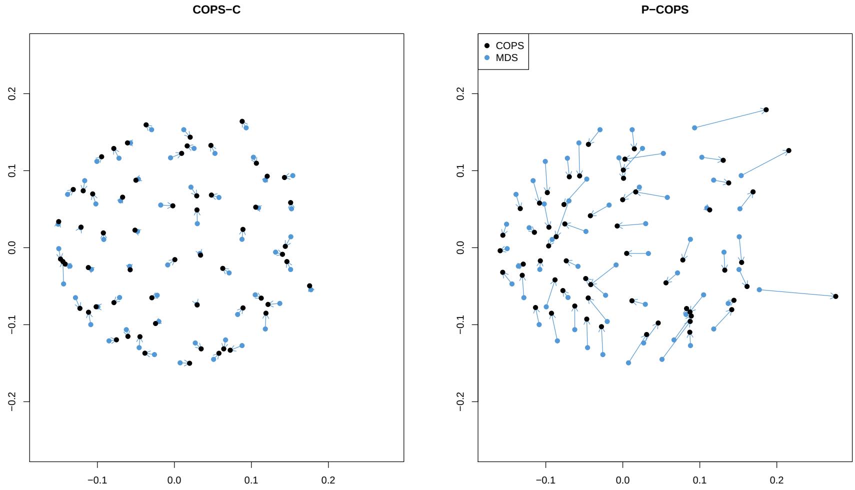
- P-COPS: `cops(dis,'P-COPS',loss='powerstress')`

```
Call: [1] "[deleted]"

Model: COPS with powerstress loss function and parameters kappa= 1.853 lambda= 8.987 nu= 0.579

Number of objects: 60
MDS loss value: 0.07394
OPTICS cordillera: Raw 5.315 Normed 0.1217
Cluster optimized loss (coploss):  -0.3079
MDS loss weight: 1  OPTICS cordillera weight: 3.138
Number of iterations of ALJ optimization: 134
```

COPS Mental States - I

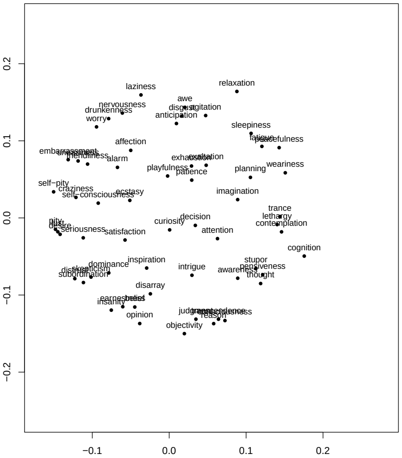


COPS Mental States - I

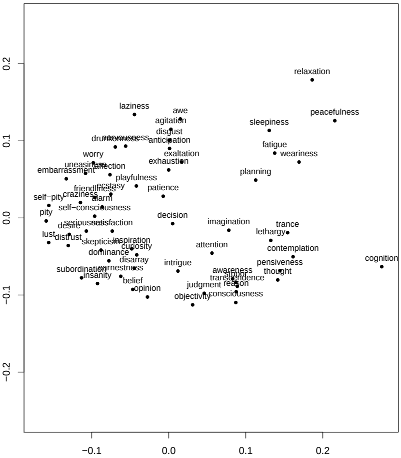


COPS Mental States - II

COPS–C



P–COPS



Why Stop with COPS?

We can go further than COPS:

- Other structures might be of interest
- Other transformations might be of interest
- Other dimensionality reduction methods might be of interest

We can rehash ideas from COPS:

- Idea behind P-COPS is rather flexible
- Conceptual and computational framework for hyperparameter selection by structure considerations
- Building blocks: θ–parametrized loss function, structuredness index(es), combination and algorithm for outer optimization.

With MDS-type losses we call this STOPS (Structure Optimized Proximity Scaling; Rusch et al., 2017).

Dimensionality Reduction - II

In MDS-type dimension reduction (proximity scaling) we have a loss function that measures misfit

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

with $\delta^*_{ij} = f_{ij}(\delta_{ij}; \theta)$ and $d^*_{ij} = g_{ij}(d_{ij}; \theta)$ which we minimize to find the configuration X given θ

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

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

STOPS - I

- We capture $p = 1, \ldots, P$ structures by indices $I_p(X(\theta); \gamma)$.
- We combine the misfit and the indices to stoploss(θ)

Two STOPS models

- Additive STOPS

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

- Multiplicative STOPS

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

v_0.. stressweight (redundant), $v_1, \ldots, v_P$... structuredness weights, γ... (optional) metaparameters for structuredness indices

STOPS - II

For hyperparameter selection we then need to find

$$\arg\min_{\vartheta} \; \mathrm{aSTOPS}(\theta, v_0, \ldots, v_k; \Delta)$$

or

$$\arg\min_{\vartheta} \; \mathrm{mSTOPS}(\theta, v_0, \ldots, v_k; \Delta)$$

where $\vartheta \subseteq \{\theta, v_0, \ldots, v_k\}$. Typically ϑ will be a subset of all possible parameters here (e.g., the weights might be given *a priori*, so $\vartheta = \theta$).

Structures and Indices - I

C-Structuredness Indices:

- They capture the essence of a particular structure in a configuration.
- They should be numerically high (low) the more (less) structure.
- They are solely a function of X (not of Δ and σ).
- They are bound from above and below, i.e., have unique finite minima and maxima.
- Reasonably regular in their behaviour as a function of the c-structuredness.
- They quantify what a human may perceive in the configuration.

Structures and Indices - II

- 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)
- 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-Functionality: Pairwise functional, smooth, noise-free relationship between axes (mean pairwise maximum edge value; Reshef et al. 2011)

Structures and Indices - III

- C-Linearity: Points lie close to linear subspace (maximal multiple correlation)
- C-Manifoldness: Points lie close to a smooth sub manifold (maximal correlation; Sarmanov, 1958)
- C-Nonmonotonicity: Deviation from monotonicity of axes (pairwise maximal maximum assymmetry score; Reshef et al. 2011)
- C-Ultrametric: How well is the overall distance variablity explained by an ultrametric (VAF and DAF)
- C-Randomness: How close to a random pattern (under some model) is the configuration (not clear yet)
- C-Faithfulness: How accurate is the neighbourhood of Δ^* preserved in D^* (adjusted M_d index of Chen & Buja, 2013)

Any other ideas?

Optimization-I

We need to find

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

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

Optimization-II

The difficulty lies in how to optimize over ϑ

- Inner minimization is costly
- Stoploss is a hard function to optimize (we basically only know function evaluations)
- Estimation of Step 1 may be noisy (premature termination, local minimum)
- We need a way to solve step 2 with a global optimization
  - only knowing target function values at some parameters
  - as little function evaluations as possible
  - the possibility that the function evaluations are noisy

Optimization-III

This can be done with Efficient Global Optimization (Bayesian Optimization).

- Black box global optimization if target function is costly
- The surrogate model allows to deal with noise
- Works well in low dimensions

Strategy is popular for hyperparameter tuning in machine learning

Optimization-IV

The idea behind this approach

- Choose a (flexible) surrogate model (prior)
- Evaluate the target function at some values (data)
- Update the prior with the function evaluations (posterior)
- Maximize an acquisition function (e.g., expected improvement (EI)) over the posterior surface
- Maximal EI suggests a candidate parameter combination
- Evaluate at candidate and repeat

One samples the "best" candidate point given the current knowledge and model.

Optimization-V

We use two types of priors:

- Simple Kriging model (Gaussian Process) with covariance kernels (Roustant et al., 2012)
  - Squared Exponential (“Gaussian”; very smooth)
  - Matern 5/2 and 3/2 (smooth)
  - Exponential (Ohrnstein Uhlenbeck process; very rough)
  - Power exponential (rough, but less so then OU)
  - Appears good for inner optimization by gradient methods or SVD
- Treed Gaussian Process with Jumps to Linear Models (Grammacy, 2007)
  - Nonstationary process by partitioning
  - Allows flexible combination of different GP, piecewise linear trends, jumps
  - Appears good for inner part estimated with majorization

R Package stops

All of this is implemented in the R package stops

- High level function for COPS `cops(delta,variant,...)`
- High level function for STOPS `stops(delta,loss,...)`
- Prespecified MDS models (argument `loss`) for STOPS and P-COPS 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 for STOPS also are Isomap, t-SNE, Diffusion Map
- Optimization with Bayesian optimization (`kriging`, `tgp`) or `ALJ` or simulated annealing (`SANN`) or a particle swarm algorithm (`pso`).
- Features various structuredness indices
- S3 methods: `plot, summary, print, coef, residuals, plot3d, plot3dstatic`

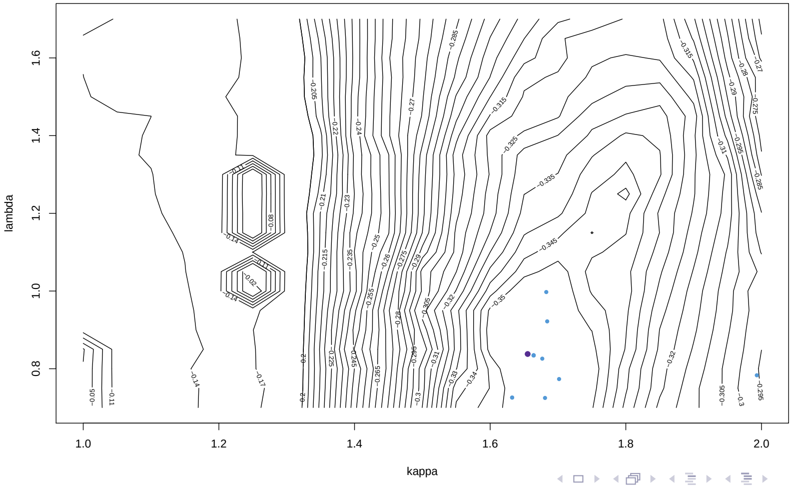
Example: Mental States - I

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

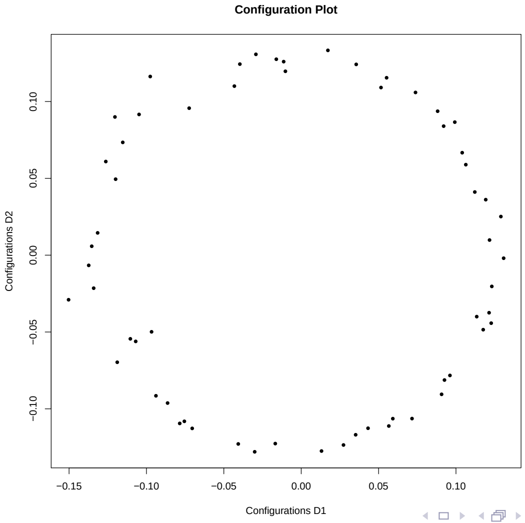
```
R> res1 <- stops(dis,loss="powermds",theta=c(1,1,1),structures=c("ccluste
R> res1
```

```
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
```

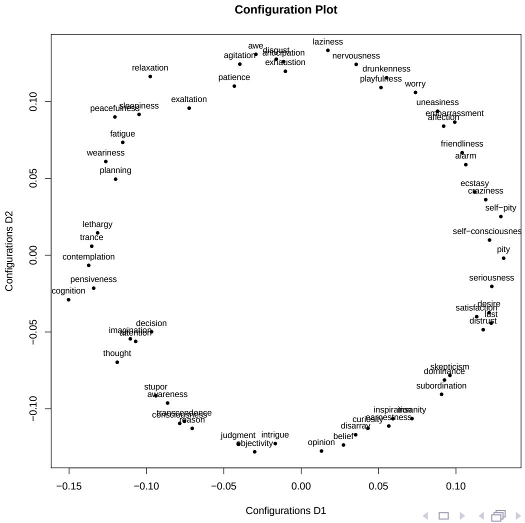
Example: Mental States - III



Example: Mental States - IV



Example: Mental States - IV



Summary

COPS

- We presented a new dimension reduction technique to obtain clustered configurations: COPS
- Two versions (COPS-C and P-COPS)

STOPS

- A framework for hyperparameter optimization in MDS based on structure considerations
- Generalization of P-COPS

Outlook

For STOPS

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

Beyond that

- We are working on a general framework for directly obtaining structured configurations by penalization
- Very much at the beginning

References

- Ankerst, M., Breunig, M., Kriegel, H.-P. & Sander, J. (1999) OPTICS: Ordering points to identify the clustering structure, ACM Sigmod Record 28, 49–60.
- 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.
- Luus, R. & Jaakola, T. (1973) Optimization by direct search and systematic reduction of the size of search region, AIChE Journal, 19, 760–766.
- 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. (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. (2015) COPS: Cluster optimized proximity scaling. Report 2015/1, Discussion Paper Series / Center for Empirical Research Methods, 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.

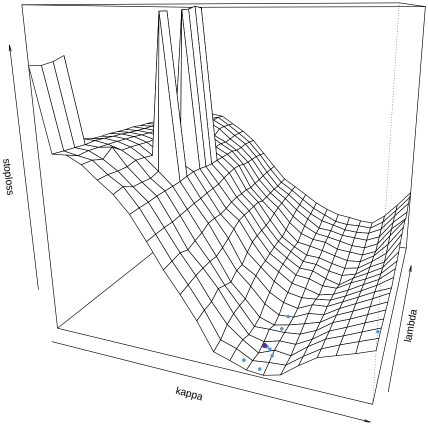
Backup Slides

Optimization Details

Adaptive Luus-Jakola Algorithm (ALJ): An adaptation of Luus-Jakola search (Luus & Jaakola, 1973)

- Sample $\theta^{(i)}$ from within t-orthotope $[l, u]^t$ with l, u are lower, upper boundaries
- Set d to be the length of the search space
- Repeat until termination (accd, maxiter, acc) :
  - Pick $a^{(i)} \sim U_t(-d, d)$
  - Set $\theta^{(i+1)} \leftarrow \theta^{(i)} + a^{(i)}$
  - If coploss$(\theta^{(i+1)}) <$ coploss$(\theta^{(i)})$ set $\theta^{(opt)} = \theta^{(i+1)}$, else set $d = d \cdot s$
- Here (this is the customized part): $s = o \cdot \frac{m+1-i}{m}$, $m = \min\left(\left\lfloor \frac{\log(accd) - \log(\max(u-l))}{\log(o)} \right\rfloor, maxiter\right)$ and $0 \leq o \leq 1$.

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