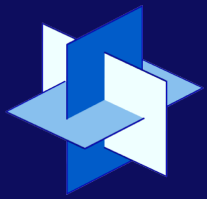
Results of Cluster Analysis in the Occasion of the Autumn Meeting of AG DANK

Marcus Weber, Susanna Kube*

Zuse Institute Berlin

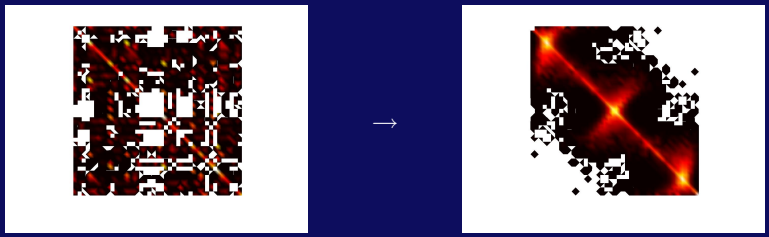
DFG Research Center
"Matheon"



What does PCCA+ do?

Row stochastic matrix T:

- $T(i,j) =$ Transition probability $i \to j$ in a detailed balanced Markov chain.
- $T(i,j) = \frac{S(i,j)}{\sum_l S(i,l)}$, where S is a symmetric measure of similarity.



Eigenvector Transformation

T_1	0	0
0	T_2	0
0	0	T_3

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2.02 & -0.55 \\ 1 & -2.02 & -0.55 \\ 1 & 0.48 & -0.91 \\ 1 & 0.48 & -0.91 \\ 1 & 0.50 & 1.24 \\ 1 & 0.50 & 1.24 \end{pmatrix} \cdot \begin{pmatrix} 0.20 & 0.41 & 0.39 \\ -0.40 & 0.33 & 0.07 \\ 0.00 & -0.47 & 0.47 \end{pmatrix}$$

$\tilde{T}_1$	E_{12}	E_{13}
E_{32}	$\tilde{T}_2$	E_{23}
E_{31}	E_{32}	$\tilde{T}_3$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0.8 & 0.2 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2.02 & -0.55 \\ 1 & -1.52 & -0.62 \\ 1 & 0.23 & -0.66 \\ 1 & 0.48 & -0.91 \\ 1 & 0.50 & 1.03 \\ 1 & 0.50 & 1.24 \end{pmatrix} \cdot \begin{pmatrix} 0.20 & 0.41 & 0.39 \\ -0.40 & 0.33 & 0.07 \\ 0.00 & -0.47 & 0.47 \end{pmatrix}$$

$$\boxed{\chi = X\mathcal{A}}$$

Uniqueness of clustering

Theorem

Let

i) $\sum_{i=1}^{k} \widetilde{\chi}_i = e$,

ii) for all $i = 1, \ldots, k$ and $l = 1, \ldots, N$: $\widetilde{\chi}_i(l) \geq 0$,

iii) $\widetilde{\chi} = \widetilde{X}\widetilde{\mathcal{A}}$ with $\widetilde{\mathcal{A}}$ regular,

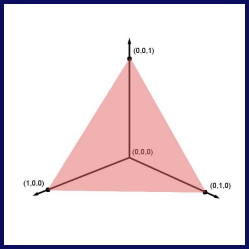
iv) for all $i = 1, \ldots, k$ there exists $l \in \{1, \ldots, N\}$ with $\widetilde{\chi}_i(l) = 1$.

Then

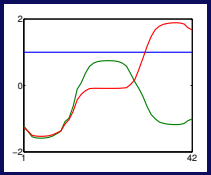
- 3 out of 4: easy to assure
- all 4: unique solution for almost characteristic functions

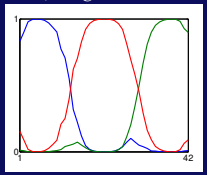
Inner Simplex Algorithm

- consider the n rows of $X \in \mathbb{R}^n$ as points in $\mathbb{R}^{N_C}$
- find those points which are the corners of a simplex
- these points define the transformation
- Deuflhard, Weber (2003)



Example n-butane: $N = 42$ states, $N_C = 3$ cluster





Quality Measurement

What is the appropriate number of clusters?

- minChi-indicator: the positivity requirement should be satisfied as good as possible

$$\text{minChi} = |\min_{i,j}(\chi(i,j))| \quad \rightarrow \quad \min!$$

- gap in the eigenvalue spectrum
- sharpness of the membership vectors χ

Computing Similarity Matrices – Outline

1. Static Clustering

consider clusters as clouds of data points which are geometrically separated from other points

$$\boxed{w(i,j) = \exp(-\beta \ \mathrm{dist}^2(i,j)), \quad i,j = 1,\ldots,N}$$

- β: parameter that corresponds to the granularity of the data
- $\mathrm{dist}(i,j) = ||p_i - p_j||_2$

$$\boxed{T = D^{-1}W}$$

$D = \mathrm{diag}(d_i)$ with $d_i = \sum_j w(i,j)$

Computing Similarity Matrices – Outline

2. Dynamic Clustering

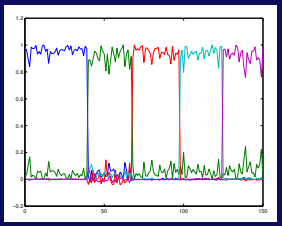
- consider data points as states in $\mathbb{R}^N$
- generate a Markov chain that jumps between these states according to the density of the points by a Metropolis-Hastings algorithm
  - start at a point p_i
  - repeat k times
    - determine its n nearest neighbors and the mean distance to these points
    - select one of the neighbors randomly $\rightarrow \quad p_j$
    - accept this new point with probability

$$\exp(-\beta(\ \mathrm{dist}(i,j) - \mathrm{meandist}(i)))$$

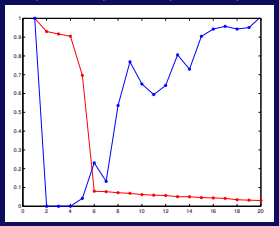
    - use p_j as new starting point p_i
  - let p_k be the final point; determine its n nearest neighbors $\{p_{k_1}, \ldots, p_{k_n}\}$
  - add 1 to the entries $w(i, k_l), \quad l = 1, \ldots n$
- repeat the above procedure for every point p_i for a fixed number of times

Results

dankdata1

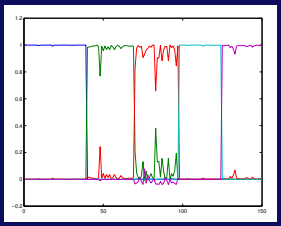


$\{38.49, 32.54, 28.49, 26.53, 23.96\}$

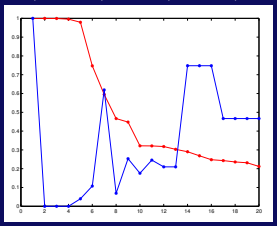


$\beta = 0.01$

dankdata4

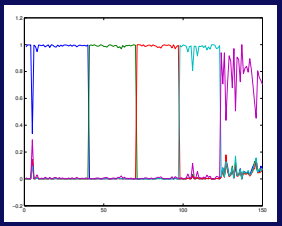


$\{38.98, 31.29, 27.40, 27.00, 25.33\}$

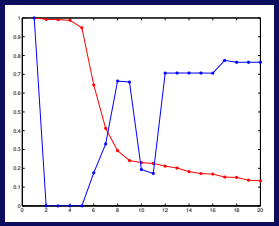


$\beta = 0.05$

dankdata6

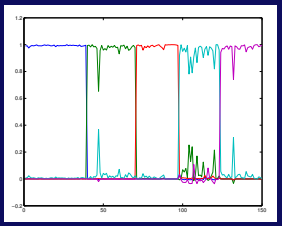


$\{40.79, 31.50, 28.41, 27.19, 22.11\}$

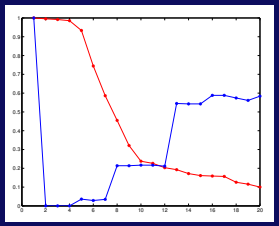


$\beta = 0.03$

dankdata9



$\{38.74, 31.56, 26.89, 26.73, 26.08\}$

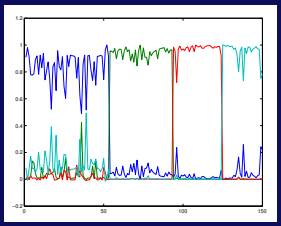


$\beta = 0.03$

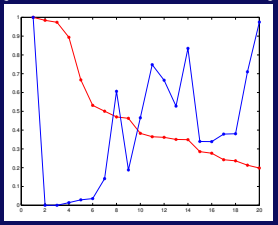
minChi allowed more (single points)

Results: Difficulties

dankdata2

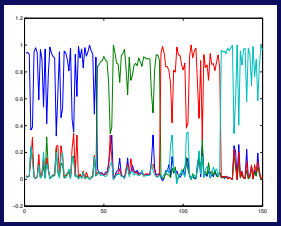


$\{48.37, 39.83, 31.19, 30.62\}$

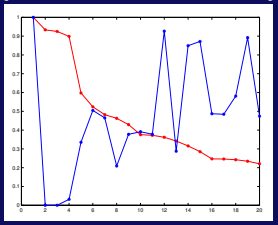


$\beta = 0.001$

dankdata5

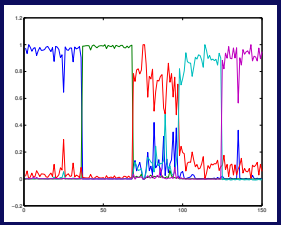


$\{42.61, 38.75, 38.34, 30.30\}$

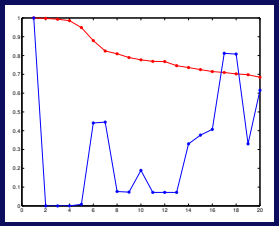


$\beta = 0.05$

dankdata8

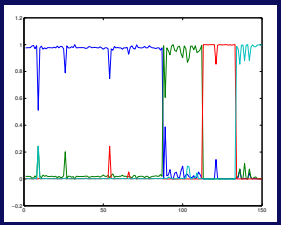


$\{38.48, 32.08, 28.45, 26.85, 24.14\}$

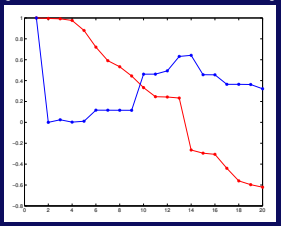


$\beta = 0.3$

dankdata11



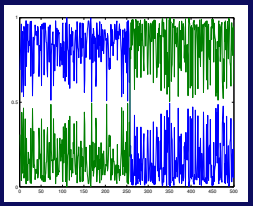
$\{85.51, 25.97, 21.30, 17.22\}$



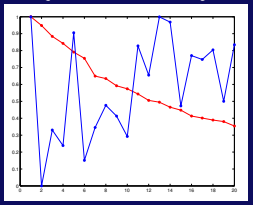
dynamic clustering:
$\beta = 1$, 10 time steps, 3 neighbors, 50 iter

Results: Cluster or not?

dankdata3



$\{252.51, 247.49\}$

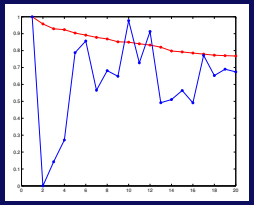


$\beta = 1$

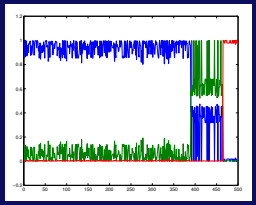
"horseshoe"

dankdata7

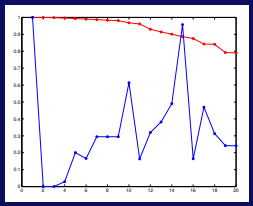
only one cluster

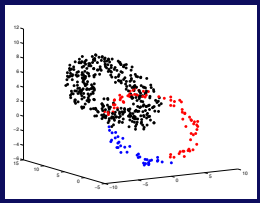


dankdata10



$\{390.42, 72.70, 36.88\}$



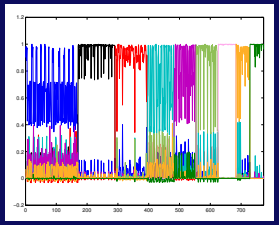


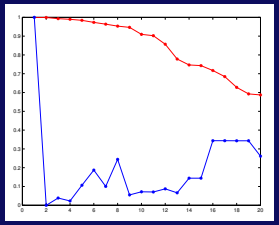
- dynamic clustering: $\beta = 20$, 10 time steps, 7 neighbors, 10 iter

Results: Praxis Test

dankdata12:

- weights: $\{127.2, 125.9, 113.9, 89.9, 83.0, 75.9, 57.7, 54.7, 44.8\}$
- dynamic clustering: $\beta = 1$, 5 time steps, 20 neighbors, 20 iterations
- gap in the spectrum as indicator

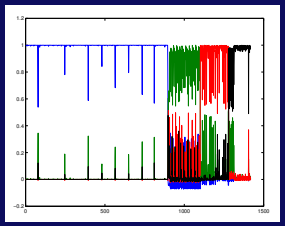


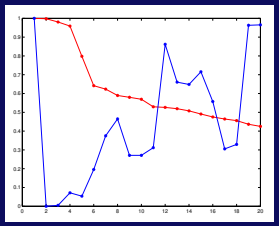


Results: Praxis Test

dankdata13:

- weights: $\{889.8, 196.6, 190.7, 140.9\}$
- $\beta = 0.01$
- sharpness of χ as criterion





Conclusions

- Robust Perron Cluster Analysis finds a hidden block-diagonal structure in a stochastic matrix.
- it does not only find clusters, but it also provides information about transition states (soft clustering method)
- static clustering:
  - number of clusters strongly depends on the distance measure
  - we can only find clusters which are separated w.r.t. this distance measure

Thank you for your attention!!!