Clustering Matrices: A Metric Learning Approach to Disease Subtyping in Mental Health

Hanchao Zhang

Division of Biostatistics, Department of Population Health,
Grossman School of Medicine, New York University

March 21, 2023

Acknowledgements

Collaborators

This talk presents joint work with the following:

- Thaddeus Tarpey (Grossman School of Medicine, NYU)
- Emily R. Stern (Grossman School of Medicine, NYU)
- Alessandro S. De Nadai (Harvard Medical School)

Fundings

This work is supported by NIMH grants:

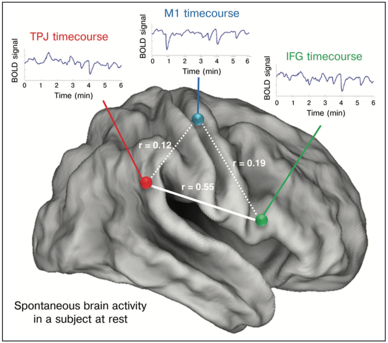
- R01 MH099003
- R01 MH126981
- R01 MH111794
- R01 MH111794
- R61/R33 MH107589

Outline for section 1

Functional Connectivity Matrix

Functional Connectivity

Functional connectivity is defined as the temporal coincidence of spatially distant neurophysiological events.



Functional Connectivity (Gillebert and Mantini, 2013)

Functional Connectivity **Matrix**

For each participant i, let $\boldsymbol{y}_{ij} \in \mathbb{R}^T$ be the longitudinal measurement of blood oxygen level-dependent (BOLD) signal on the region of interest j, $j = 1, 2, \ldots, p$.

The functional connectivity matrix for participant i: $\boldsymbol{\Sigma}_i = \mathbf{Cov}(\boldsymbol{y}_i) \succeq \mathbf{0}$

Self-Consistency Clustering Algorithms

Scalar Outcomes

K-Means algorithm (Steinhaus et al., 1956)

$$\underset{\mathcal{C}}{\text{minimize}} \underbrace{\frac{1}{n}\sum_{i=1}^{m}\sum_{k\in\mathcal{C}_i}\left\|x_k - \bar{x}_{\mathcal{C}_i}\right\|^2}_{\text{within cluster sum of squares}} \quad \textbf{OR} \quad \underset{\mathcal{C}}{\text{maximize}} \underbrace{\sum_{i=m}\frac{n_{\mathcal{C}_i}}{n}\cdot\left\|\bar{x}_{\mathcal{C}_i}\right\|^2}_{\text{between cluster sum of squares}}$$

Functional Outcomes

Clustering Functional data (Tarpey and Kinateder, 2003)
$\boldsymbol{y}_i(t), i = 1, \ldots, n, t \in T$, typically a compact real interval, $y_i(t) =$ function

$$\boldsymbol{y}_i(t) = \boldsymbol{b}'(t)\ \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i(t) = \sum_{j=1}^{\infty}\beta_{ij}b_j(t) + \epsilon_i(t).$$

- $\boldsymbol{b} = (b_1(t), \ldots, b_p(t), \ldots)'$ is basis functions
- $\boldsymbol{\beta}_i = (\beta_{1i}, \ldots, \beta_{ip}, \ldots)'$ is a vector of basis coefficients

Perform K-Means or other algorithms on basis coefficients $\boldsymbol{\beta}_i$

Positive Semi-Definite Matrix Outcomes

Consider that for each observation i, $i = 1, 2, \ldots, n$, we observe p functional outcomes $\boldsymbol{y}_{ij}(t)$ with mean 0, $j = 1, 2, \ldots, p$. Then we can obtain a positive semi-definite matrix for subject i:

$$\boldsymbol{\Psi}_i = \int_T \boldsymbol{y}_i(t)^T \boldsymbol{y}_i(t)\ dt, \boldsymbol{\Psi}_i \succcurlyeq 0,$$

where $\boldsymbol{\Psi}_i \succcurlyeq 0$ means $\boldsymbol{\Psi}_i$ is positive semi-definite matrix. (All the eigenvalues of $\boldsymbol{\Psi}_i$ are larger and equal to 0).

Clustering Algorithm Approaches

- cluster subjects by $\boldsymbol{\Psi}_i$'s, $i = 1, 2, \ldots, n$
- vectorize $\boldsymbol{\Psi}_i$'s and treat it as vector
- consider some distance metrics for matrix similarity
- consider the probability distribution (e.g., Wishart distribution)

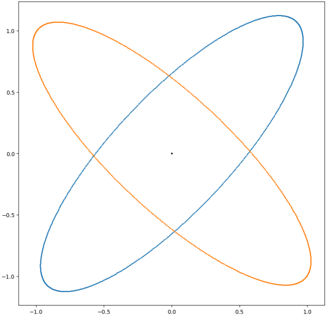
Outline for section 2

Distance Metrics for Matrices

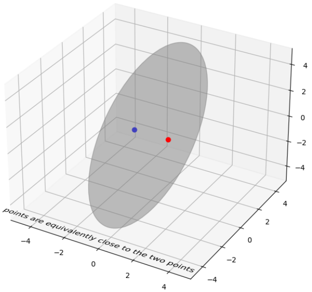
Euclidean Distance (chapter 2 Minh and Murino, 2017)

let $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2 \in \mathbb{R}^{2\times 2}$ be two positive semi-definite matrices. The Euclidean distance between two matrices $d_E(\mathbf{\Psi}_1, \mathbf{\Psi}_2)$ can be represented by points in $\mathbb{R}^3$

$$d_E(\mathbf{\Psi}_1, \mathbf{\Psi}_2) = \|\mathbf{\Psi}_1 - \mathbf{\Psi}_2\|_F^2 = \|\mathrm{vec}(\mathbf{\Psi}_1^T) - \mathrm{vec}(\mathbf{\Psi}_2^T)\|^2$$

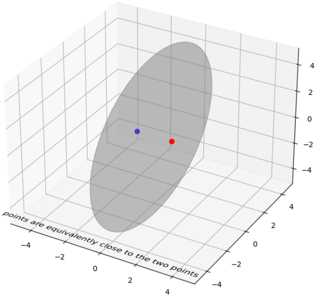


2 positive semi-definite matrices

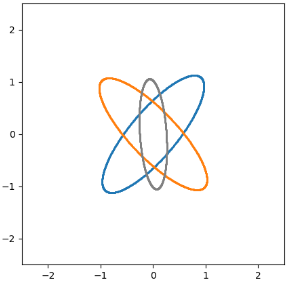


the vectorized matrices

Distance Metrics for Matrices



the vectorized matrices



Disadvantage of Euclidean Distance

- matrices with similar shapes are clustered into different groups

Distance Metrics for Matrices (chapter 2 Malhi et al., 2017)

Other Metrics

Affine-invariant Riemannian Distance

$$d_{aiE}(\mathbf{A}, \mathbf{B}) = || \log(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}})||_F$$

Log-Determinant Divergences

$$d^1_{\log\det}(\mathbf{A}, \mathbf{B}) = \mathbf{tr}(\mathbf{B}^{-1}\mathbf{A} - \mathbf{I}) - \log\det(\mathbf{B}^{-1}\mathbf{A})$$

Symmetric Stein Divergence

$$d^2_{\text{stein}}(\mathbf{A}, \mathbf{B}) = \log\det(\frac{\mathbf{A}+\mathbf{B}}{2}) - \frac{1}{2}\log\det(\mathbf{AB})$$

Disadvantages

- does not consider the structure (shape) of p.s.d matrices

Outline for section 3

Common Principal Components (CPC) (Flury, 1984)

Definition

Let $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n$ be positive definite symmetric matrix of dimension $p \times p$, we wish to find a orthonormal matrix $\boldsymbol{B}$ which makes the $\boldsymbol{\Psi}_i$'s simultaneously "as diagonal as possible":

Objective Function

Let $\boldsymbol{F}_i$ be the transformed $\boldsymbol{\Psi}_i$ by $\boldsymbol{B}$:

$$\boldsymbol{F}_i = \boldsymbol{B}^T \boldsymbol{\Psi}_i \boldsymbol{B}$$

To make sure that $\boldsymbol{F}_i$'s, $i = 1, \ldots, n$ are as diagonal as possible, we wish to minimize:

$$\underset{\boldsymbol{B}}{\textbf{minimize}} \quad \prod_{i=1}^{n} \left\{ \frac{\det\left(\textbf{diag}(\boldsymbol{F}_i)\right)}{\det\left(\boldsymbol{F}_i\right)} \right\} = \prod_{i=1}^{n} \left\{ \frac{\det\left(\textbf{diag}(\boldsymbol{B}^T \boldsymbol{\Psi}_i \boldsymbol{B})\right)}{\det\left(\boldsymbol{B}^T \boldsymbol{\Psi}_i \boldsymbol{B}\right)} \right\},$$

where $\det(F_i) \leq \det(\textbf{diag}(F_i))$.

Algorithms

- FG-algorithm (Flury and Constantine, 1985)
- MM algorithms (Browne and McNicholas, 2014)
- R algorithm (Hallin et al., 2014)

Unconstrained Common Principal Components for Clustering Matrices

Stiefel Manifold

the Stiefel manifold $V_k(\mathbb{R}^p)$ is the set of all orthonormal k-frames in $\mathbb{R}^p$

$$V_k(\mathbb{R}^p) = \{\boldsymbol{A} \in \mathbb{R}^{n\times p} : \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}_p\}$$

Self-Consistency Algorithm Based on CPC

Let $S \subset \mathbb{R}^{p\times p}, S \succeq 0$ denote a set of p.s.d. matrices. For each $\boldsymbol{B} \in V_p(\mathbb{R}^p)$, define:

$$D_{\boldsymbol{B}}(S) = \Big\{\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}, \boldsymbol{\Psi} \succeq 0 : \|\boldsymbol{\Psi} - \underbrace{\boldsymbol{B}\ \mathbf{diag}(\boldsymbol{F_B})\ \boldsymbol{B}^T}_{\hat{\boldsymbol{\Psi}}}\|_F \ \le \|\boldsymbol{\Psi} - \boldsymbol{A}\ \mathbf{diag}(\boldsymbol{F_A})\ \boldsymbol{A}^T\|_F, \quad \boldsymbol{B} \neq \boldsymbol{A}, \boldsymbol{A} \in V_p(\mathbb{R}^p)\Big\}.$$

Therefore, each matrix in set $D_{\boldsymbol{B}}(S)$ shares common principal components $\boldsymbol{B}$ that can make them "as diagonal as possible".

Note

Since we have $\boldsymbol{\Psi} = \boldsymbol{B}\boldsymbol{F_B}\boldsymbol{B}^T = \boldsymbol{A}\boldsymbol{F_A}\boldsymbol{A}^T$, we can redefine $D_{\boldsymbol{B}}(S)$ as follow:

$$D_{\boldsymbol{B}}(S) = \Big\{\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}, \boldsymbol{\Psi} \succeq 0 : \|\boldsymbol{F_B} - \mathbf{diag}(\boldsymbol{F_B})\|_F \ \le \|\boldsymbol{F_A} - \mathbf{diag}(\boldsymbol{F_A})\|_F, \quad \boldsymbol{B} \neq \boldsymbol{A}, \boldsymbol{A} \in V_p(\mathbb{R}^p)\Big\}.$$

Unconstrained Common Principal Components for Clustering Matrices

Self-Consistency Algorithm Based on CPC

Let $S \in \mathbb{R}^{p \times p}, S \succeq 0$ denote a measurable set. For each $\boldsymbol{B} \in V_p(\mathbb{R}^p)$, define:

$$D_{\boldsymbol{B}}(S) = \{\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}, \boldsymbol{\Psi} \succeq 0 : \|\boldsymbol{\Psi} - \boldsymbol{B} \text{ diag}(\boldsymbol{F_B}) \ \boldsymbol{B}^T\|_F \ \leq \|\boldsymbol{\Psi} - \boldsymbol{A} \text{ diag}(\boldsymbol{F_A}) \ \boldsymbol{A}^T\|_F, \boldsymbol{B} \neq \boldsymbol{A}, \boldsymbol{A} \in V_p(\mathbb{R}^p)\}.$$

Unconstrained CPC for Matrices Clustering

Algorithm **Clustering Matrices Using Unconstrained CPC**

Start with an initial partition of all matrices into K clusters

1: for each cluster k, $k = 1, 2, \ldots, K$, estimate the common principal component $\boldsymbol{B}_k$.

2: assign individual matrices $\boldsymbol{\Psi}_i$ to cluster k if

$$k^* = \underset{k=1,\ldots,K}{\arg\min} \|\boldsymbol{\Psi}_i - \boldsymbol{B}_k \text{diag}(\boldsymbol{F}_{\boldsymbol{B}_k}) \boldsymbol{B}_k^T\|_F$$

repeat steps 1 and 2 until convergence.

Partial Common Principal Components

Self-Consistency Algorithm Based on Partial CPC

Let $S \in \mathbb{R}^{p\times p}, S \succeq 0$ denote a measurable set. For each $\boldsymbol{B} := \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m\right) \in V_m(\mathbb{R}^p)$,

$$D_{\boldsymbol{B}}(S) = \Big\{\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}, \boldsymbol{\Psi} \succeq 0 : \|\boldsymbol{\Psi} - \sum_{r=1}^{m} f_{\boldsymbol{B}_r}\boldsymbol{\beta}_r\boldsymbol{\beta}_r^T\|_F \le \|\boldsymbol{\Psi} - f_{\boldsymbol{A}_r}\boldsymbol{\alpha}_r\boldsymbol{\alpha}_r^T\|_F,$$
$$\boldsymbol{B} \neq \boldsymbol{A}, \boldsymbol{A} \in V_m(\mathbb{R}^p)\Big\},$$

where $f_1, \ldots, f_p$ are the the diagonal elements of $\boldsymbol{F}$, and $m \le p$.

Unconstrained CPC for Matrices Clustering

Algorithm **Clustering Matrices Using Unconstrained CPC**

Start with an initial partition of all matrices into K clusters

1: for each cluster k, $k = 1, 2, \ldots, K$, estimate the common principal component $\boldsymbol{B}_k$.

2: assign individual matrices $\boldsymbol{\Psi}_i$ to cluster k if $k^* = \underset{k=1,\ldots,K}{\arg\min} \|\boldsymbol{\Psi}_i - \sum_{r=1}^{m} f_{\boldsymbol{B}_{k_r}}\boldsymbol{\beta}_r\boldsymbol{\beta}_r^T\|_F$

repeat steps 1 and 2 until convergence.

Outline for section 4

Simulation

Simulation Settings

let $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ be the two common eigenvectors for the two clusters

$$\boldsymbol{B}_1 = \begin{pmatrix} \cos(\beta_1) & -\sin(\beta_1) \\ \sin(\beta_1) & \cos(\beta_1) \end{pmatrix} \qquad \boldsymbol{B}_2 = \begin{pmatrix} \cos(\beta_2) & -\sin(\beta_2) \\ \sin(\beta_2) & \cos(\beta_2) \end{pmatrix},$$

where β_1 and β_2 be 2 scalars from 0 to 2π. Let $|\beta_1 - \beta_2| = \theta$ be the differences between two eigenvectors.

let $\boldsymbol{\lambda}_{\boldsymbol{B}_1 i} = [\lambda_{\boldsymbol{B}_1 i1}, \lambda_{\boldsymbol{B}_1 i2}]$ and $\boldsymbol{\lambda}_{\boldsymbol{B}_2 i} = [\lambda_{\boldsymbol{B}_2 i1}, \lambda_{\boldsymbol{B}_2 i2}]$ be the eigenvalues for the two clusters. Denote $\boldsymbol{\Lambda}_{\boldsymbol{B}_1 i} = \text{diag}(\boldsymbol{\lambda}_{\boldsymbol{B}_1 i})$, and $\boldsymbol{\Lambda}_{\boldsymbol{B}_2 i} = \text{diag}(\boldsymbol{\lambda}_{\boldsymbol{B}_2 i})$, where $\lambda \sim \chi^2(\text{df})$.

Then we can obtain our simulated matrices:

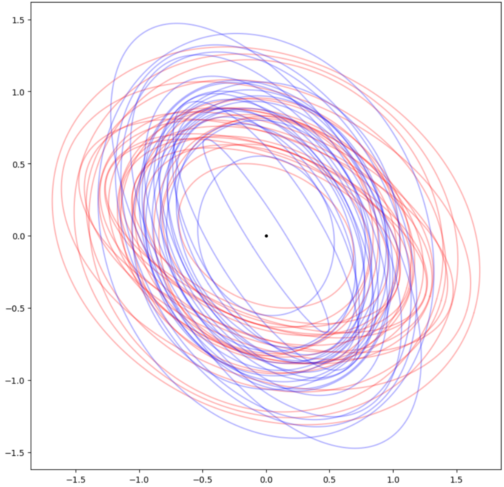
$$\boldsymbol{\Psi}_{1i} = \boldsymbol{B}_1 \boldsymbol{\Lambda}_{\boldsymbol{B}_1 i} \boldsymbol{B}_1^T + \boldsymbol{E}_1 \qquad \boldsymbol{\Psi}_{2i} = \boldsymbol{B}_2 \boldsymbol{\Lambda}_{\boldsymbol{B}_2 i} \boldsymbol{B}_2^T + \boldsymbol{E}_2,$$

where $\boldsymbol{E}_1$, and $\boldsymbol{E}_1$ are random error with mean 0.

Note

- θ denotes how close the two clusters of matrices are
- $\boldsymbol{E}$ is some random perturbation on eigenvector and eigenvalues.

Simulations



$\theta = \pi/5$, $df = 5$

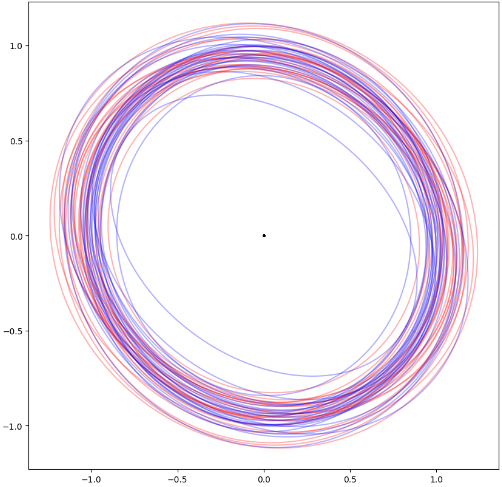
Simulation

- $\theta = \pi/5$
- eigenvalues $\sim \chi^2(5)$

Classification Error

- CPCA $= 0$
- rCPCA $= 0$
- Euclidean Distance $= 0.28$
- Affine Invariance Divergence $= 0.34$
- Log-Determinant Divergence $= 0.38$

Simulations



$\theta = \pi/5$, $df = 5$

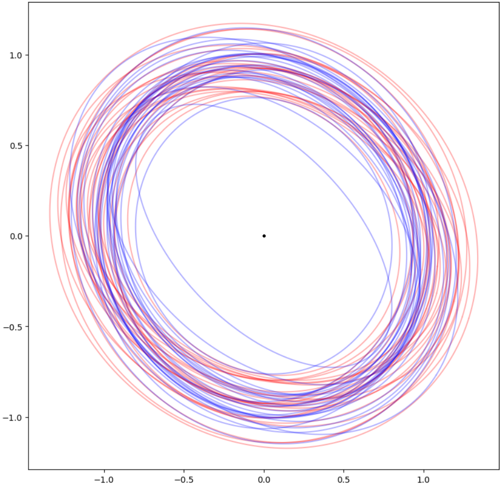
Simulation

- $\theta = \pi/15$
- eigenvalues $\sim \chi^2(40)$

Classification Error

- CPCA $= 0$
- rCPCA $= 0.02$
- Euclidean Distance $= 0.4$
- Affine Invariance Divergence $= 0.4$
- Log-Determinant Divergence $= 0.44$

Simulations



$\theta = \pi/5$, $df = 5$

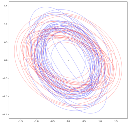
Simulation

- $\theta = \pi/15$
- eigenvalues $\sim \chi^2(20)$
- noise $= 5\%$

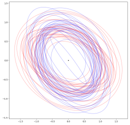
Classification Error

- CPCA $= 0.02$
- rCPCA $= 0.02$
- Euclidean Distance $= 0.4$
- Affine Invariance Divergence $= 0.4$
- Log-Determinant Divergence $= 0.44$

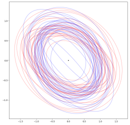
Simulations



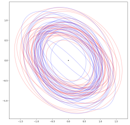
$\theta = \pi/5$, $df = 5$



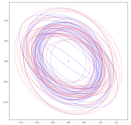
$\theta = \pi/6$, $df = 5$



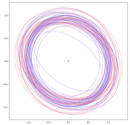
$\theta = \pi/8$, $df = 5$



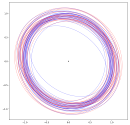
$\theta = \pi/10$, $df = 5$



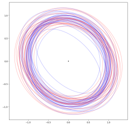
$\theta = \pi/15$, $df = 5$



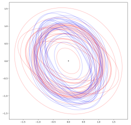
$\theta = \pi/15$, $df = 20$



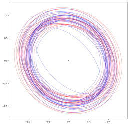
$\theta = \pi/15$, $df = 40$



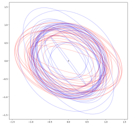
$\theta = \pi/15$, $df = 20$, noise $= 1\%$



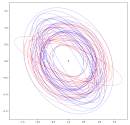
$\theta = \pi/15$, $df = 20$, noise $= 2\%$



$\theta = \pi/15$, $df = 20$, noise $= 4\%$



$\theta = \pi/15$, $df = 20$, noise $= 6\%$



$\theta = \pi/15$, $df = 20$, noise $= 10\%$

Simulations

Simulation Results

Classification Errors					
Methods	$\theta = \pi/5$ $df = 5$	$\theta = \pi/6$ $df = 5$	$\theta = \pi/8$ $df = 5$	$\theta = \pi/10$ $df = 5$	$\theta = \pi/15$ $df = 5$
CPCA	0.00	0.00	0.00	0.00	0.00
rCPCA	0.00	0.00	0.02	0.00	0.02
Frobenius	0.28	0.32	0.32	0.34	0.34
Aff. Div.	0.34	0.38	0.42	0.42	0.44
Log-Det	0.38	0.38	0.44	0.44	0.46
Methods	$\theta = \pi/15$ $df = 20$	$\theta = \pi/15$ $df = 40$	$\theta = \pi/15$ $df = 20$ noise $= 0.01$	$\theta = \pi/15$ $df = 20$ noise $= 0.06$	$\theta = \pi/15$ $df = 20$ noise $= 0.10$
CPCA	0.00	0.00	0.02	0.10	0.18
rCPCA	0.10	0.12	0.02	0.14	0.20
Frobenius	0.40	0.40	0.40	0.36	0.50
Aff. Div.	0.40	0.40	0.40	0.40	0.46
Log-Det	0.44	0.44	0.44	0.44	0.48

Outline for section 5

Referenes

R. P. Browne and P. D. McNicholas. Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, 8:217–226, 2014.

B. N. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984.

B. N. Flury and G. Constantine. Algorithm as 211: The fg diagonalization algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(2):177–183, 1985.

C. R. Gillebert and D. Mantini. Functional connectivity in the normal and injured brain. *The Neuroscientist*, 19(5):509–522, 2013.

M. Hallin, D. Paindaveine, and T. Verdebout. Efficient r-estimation of principal and common principal components. *Journal of the American Statistical Association*, 109 (507):1071–1083, 2014.

G. S. Malhi, Y. Byrow, T. Outhred, and K. Fritz. Exclusion of overlapping symptoms in dsm-5 mixed features specifier: heuristic diagnostic and treatment implications. *CNS spectrums*, 22(2):126–133, 2017.

H. Q. Minh and V. Murino. Covariances in computer vision and machine learning. *Synthesis Lectures on Computer Vision*, 7(4):1–170, 2017.

H. Steinhaus et al. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1 (804):801, 1956.

T. Tarpey and K. K. Kinateder. Clustering functional data. *Journal of classification*, 20 (1), 2003.