Optimal Linear Transformations of Functional Data for Clustering Methods

Hanchao Zhang

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

August 12, 2021

Acknowledgements

This talk presents joint work with:
Thaddeus Tarpey (New York University)

This work is supported by grants:
R01 MH099003 from the National Institute of Mental Health (NIMH)

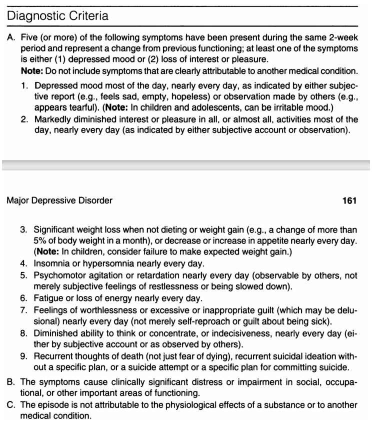
Outline for section 1

1. Motivation
   - Diagnosis of Psychiatric Illness
   - Nosology
   - Data
2. Approach – improve the clustering for functional data
   - Functional Data
   - Preconditioning - Linear Transformation
3. Reference to Other Approaches

Diagnosis of Psychiatric Illness

- Diagnostic and Statistical Manual of Mental Disorders (DSM-5)
- Nosology: the branch of medical science dealing with the classification of diseases

DSM-5 Exmaples



Diagnostic Criteria

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

Note: Do not include symptoms that are clearly attributable to another medical condition.

1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful). (**Note:** In children and adolescents, can be irritable mood.)
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).

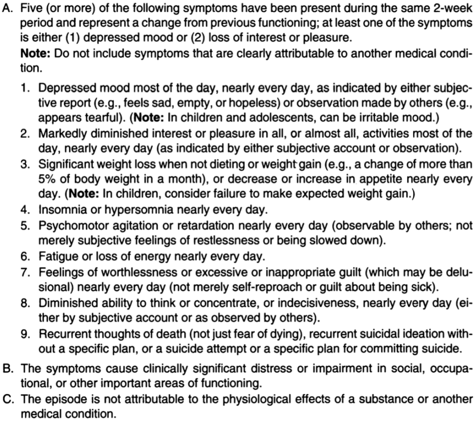
Major Depressive Disorder 161

3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. (**Note:** In children, consider failure to make expected weight gain.)
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The episode is not attributable to the physiological effects of a substance or to another medical condition.

(a) depression



A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

Note: Do not include symptoms that are clearly attributable to another medical condition.

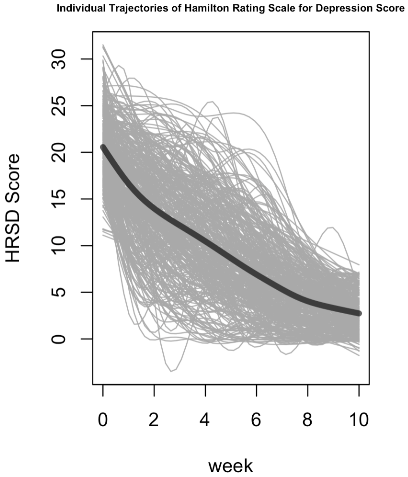
1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, or hopeless) or observation made by others (e.g., appears tearful). (**Note:** In children and adolescents, can be irritable mood.)
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. (**Note:** In children, consider failure to make expected weight gain.)
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation nearly every day (observable by others; not merely subjective feelings of restlessness or being slowed down).
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The episode is not attributable to the physiological effects of a substance or another medical condition.

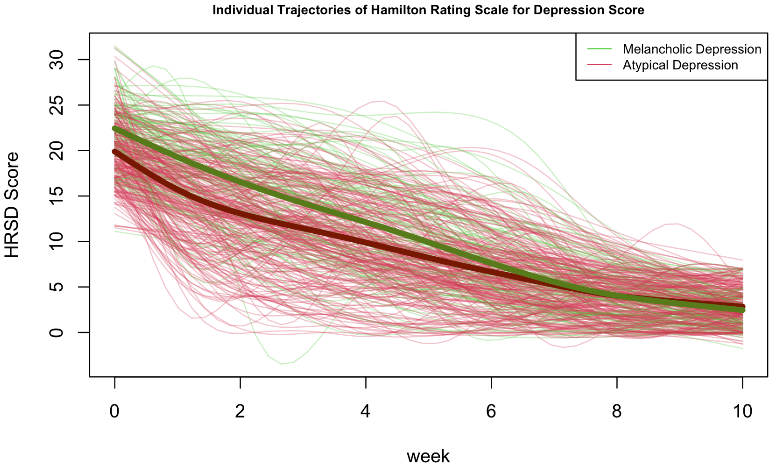
(b) bipolar

Functional Examples



- 353 patients in total
- 10 weeks open label
- all patients get drug
- melancholic depression or atypical depression diagnosed by DSM-5 criteria
- mean trajectory of the HRSD decreases

Functional Examples



Outline for section 2

1. Motivation
   - Diagnosis of Psychiatric Illness
   - Nosology
   - Data
2. Approach – improve the clustering for functional data
   - Functional Data
   - Preconditioning - Linear Transformation
3. Reference to Other Approaches

Functional Data

$\boldsymbol{y}_i(t), i = 1, \ldots, n, t \in T$, typically a compact real interval

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

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

Examples of Functional Data in Mental Health

- MRI and fMRI data (Reiss et al. 2014, *JCGS*, and Chen et al. 2014, *Biometrics*)
- EEG data of Brain (Tarpey and Petkova 2012)
- Any data that is not a point and can be seen as trajectories or expressed by basis functions

Preconditioning - Linear Transformation

The linear model

$$\boldsymbol{y}_i = \boldsymbol{X}_i \beta_i + \boldsymbol{\epsilon}_i$$

is identical to

$$\boldsymbol{y}_i = [\boldsymbol{X}_i \boldsymbol{A}^{-1}][\boldsymbol{A} \beta_i] + \boldsymbol{\epsilon}_i$$

If $\boldsymbol{A}$ is a non-sigular matrix.

The linearly transformed design matrix $\boldsymbol{X}_i A^{-1}$ can be regarded as a change in the basis representation of the functional data

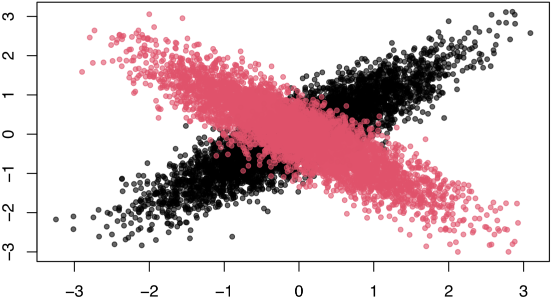
However, Clustering result based on the original coefficients β_i can differ dramatically from cluster results using $A\beta_i$

Cluster Results Changes by Linear Transformation

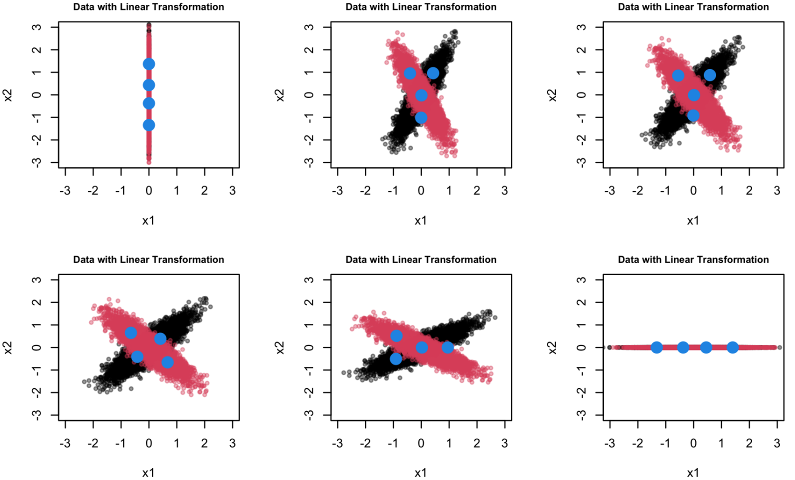
Simulated Data

Data is simulated from two bi-variate normal distributions

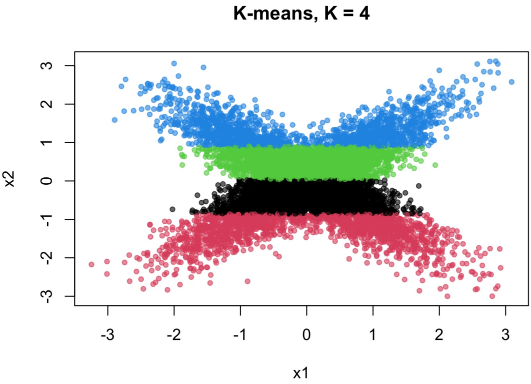
$$\boldsymbol{X}_1 \sim N(\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0.8 & -0.7\\-0.7 & 0.8\end{pmatrix}) \quad \boldsymbol{X}_2 \sim N(\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0.8 & 0.7\\0.7 & 0.8\end{pmatrix})$$



Linear Tranformation



Linear Transformation


K-means, K = 4

Goal: Find a linear transformation $\boldsymbol{A}$ to optimize the clustering

Method

Let $\boldsymbol{X}$ be the data, $\boldsymbol{A}$ be a non-singular matrix represent the linear transformation, and $\mathcal{C}$ represent the current label (e.g. diagnosis of melancholic depression or atypical depression)

$$\min L\Big(Kmeans(\boldsymbol{XA}), \mathcal{C}\Big)$$

Where

$$\boldsymbol{A} = \arg\min L\Big(Kmeans(\boldsymbol{X\tilde{A}}), \mathcal{C}\Big)$$

Question: How do we find an appropriate Loss function L?

Method

Question: The number of the distinct classes in the label might be different from the K-means (K pre-determined). Way to measure the "match-up" of the two-different clustering of a data?

Variation of Information (VI)

Given two clusterings of the same data $\mathcal{C}_1$ and $\mathcal{C}_2$, let

$$P(j,j') = \frac{\left|\mathcal{C}_j \bigcap \mathcal{C}'_j\right|}{n}, \quad j = 1, 2, \ldots, K$$

for cluster $\mathcal{C}_j$ in $\mathcal{C}_1$ and cluster $\mathcal{C}'_j$ in $\mathcal{C}_2$.

$$\text{Mutual Information} = I(\mathcal{C}_1, \mathcal{C}_2) = \sum_{j=1}^{K} \sum_{j'=1}^{K} P(j,j') \log(\frac{P(j,j')}{P(j,j), P(j',j')})$$

$$\text{Entropy for } \mathcal{C} = H(\mathcal{C}) = -\sum_{j=1}^{k} P(j) \log(P(j))$$

Variation of Information (VI) continuous

metrics to measuring similarity between two clustering of the same data

$$\text{Variation Information } = VI(\mathcal{C}_1, \mathcal{C}_2) = H(\mathcal{C}_1) + H(\mathcal{C}_2) - 2I(\mathcal{C}_1, \mathcal{C}_2)$$

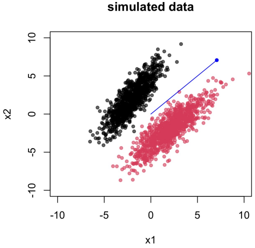
1. **Optimal:** $VI = 0$ if the two clusterings produce identical clusters (up to a re-labeling)
2. $VI > 0$ otherwise
3. $VI(\mathcal{C}_1, \mathcal{C}_2)$ is bounded by $\log(n)$ or $2\log(K^*)$, $K^* = max(K, K')$

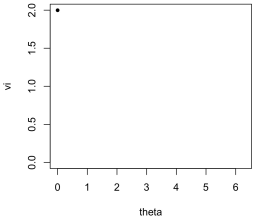
Parametrize Linear Transformation Matrix A – Two Dimension Case

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$$\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{R}$$

where $\boldsymbol{B}$ can be any unit vector in two dimensional space





Parametrize of Linear Transformation Matrix A – High Dimensional Case

$$\boldsymbol{R} = \prod_{i=2}^{p} \boldsymbol{R}_i$$

$$R_i = \begin{pmatrix} a_{11} & 0 & 0 & -b_{1i} & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ b_{i1} & 0 & 0 & a_{ii} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}_{p\times p}$$

$$a_{pq} = \cos\gamma_q \qquad b_{pq} = \sin\gamma_q$$

Obtain the Projection Matrix A – High Dimensional Case

We can find matrix A by

$$\boldsymbol{A}_{q\times p} = \boldsymbol{B}_{q\times p} \times \boldsymbol{R}_{p\times p}$$

$$\boldsymbol{A}_{q\times p} = \arg\min VI\Big(Kmeans(\boldsymbol{X}\tilde{\boldsymbol{A}}), \mathcal{C}\Big)$$

1. $\boldsymbol{B}_{q\times p}$ is an orthonormal matrix with q norm vectors in the p dimensional space (e.g. $\boldsymbol{B}_{q\times p} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}_{q\times p}$)
2. $q \le K^* - 1$

Obtain the Projection Matrix A – High Dimensional Case

Recall

- data matrix $\boldsymbol{X}_{n\times p}$
- non-singular transformation matrix $\boldsymbol{A}_{p\times q}$
- $\boldsymbol{y}_i(t) = \boldsymbol{b}'(t)\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i(t) = \sum_{j=1}^{\infty} \beta_{ij} b_j(t) + \epsilon_i(t)$

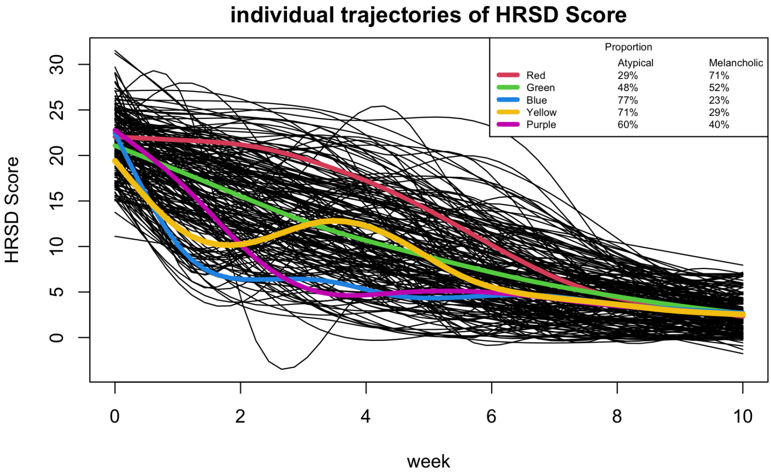
Instead of clustering the data based on the raw data $\boldsymbol{X}$, we can cluster the data based on the coefficients $\boldsymbol{\beta}_{n\times r}$.

We can obtain the transformation matrix $\boldsymbol{A}$ and cluster $\mathcal{G}$ by

$$\boldsymbol{A} = \arg\min \left[VI\left(Kmeans(\boldsymbol{\beta}\boldsymbol{A}), \mathcal{C} \right) \right]$$

$$\mathcal{G} = \arg\min_{\mathcal{G}} \sum_{i=1}^{K} \sum_{x\in i} ||\boldsymbol{\beta}\boldsymbol{A} - \mu_i||^2$$

K = 5, q = 4



	Proportion	
	Atypical	Melancholic
Red	29%	71%
Green	48%	52%
Blue	77%	23%
Yellow	71%	29%
Purple	60%	40%

Improvement on Variation Information

	K = 3	K = 4	K = 5
no transformation	2.31	2.71	2.96
q = 1	1.46	1.82	2.03
q = 2	1.40	1.78	2.02
q = 3		1.76	1.95
q = 4			1.91

Table 1: Variation Information w.r.t K and q

The variation information reduce with projection on a higher dimension for each K

Outline for section 3

1. Motivation
   - Diagnosis of Psychiatric Illness
   - Nosology
   - Data
2. Approach – improve the clustering for functional data
   - Functional Data
   - Preconditioning - Linear Transformation
3. Reference to Other Approaches

Outcome Guided K-means, L Meng, D Avram, G Tseng, Z Huo, 2020

$$\max_{C,\boldsymbol{w}} \sum_{g=1}^{G} w_g \left[\frac{BCSS_g}{TSS_g} + \lambda U_g \right]$$

$$\text{subject to} ||\boldsymbol{w}||_2 \leq 1, ||\boldsymbol{w}||_1 \leq s$$

$$U_g = 1 - \left[\frac{L(f_0)}{L(f_g)} \right]^{\frac{2}{n}}$$

where n is the number of subjects, $L(f_0)$ is the ikelihood of null model, and $L(f_g)$ is the likelihood of model f_g.
Univariate regression model f_g can be linear model, glm, cox model or other regression models.
The scale of $\frac{L(f_0)}{L(f_g)}$ and $\frac{BCSS_g}{TSS_g}$ are both $[0, 1]$

Outcome-Guided Mixture Model, Peng Liu, Yusi Fang, Zhao Ren, Lu Tang, George C. Tseng, 2021

Assume a following mixture model:

$$f(y_i; \boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_k(y_i; \boldsymbol{x})_i)$$

where $f_k(y, \boldsymbol{x})$ is the density function of cluster k. Assume $y_i | z_i = k \sim N(\beta_{0k} + \boldsymbol{\beta}^T \boldsymbol{x}_i, \sigma^2)$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{ik}(\boldsymbol{g}_i, \gamma) f(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma)$$

Reference

Peng Liu and etc. (2020), "Outcome-Guided Disease Subtyping for High-Dimensional Omics Data", arXiv:2007.11123

Lingsong Meng and etc. (2020), "Outcome-guided Sparse K-means for Disease Subtype Discovery via Integrating Phenotypic Data with High-dimensional Transcriptomic Data", arXiv:2103.09974

Jia, J. and Rohe, K. (2015), "Preconditioning the Lasso for sign consistency," Electronic Journal of Statistics, 9, 1150-1172.

Tarpey, T. (2007), "Linear Transformations and the k-Means Clustering Algorithm: Applications to Clustering Curves," The American Statistician, 61, 34–40

Meila, M. (2007), "Comparing clusterings? an information based distance," Journal of Multivariate Analysis, 98, 873-895.