Southern University of Science and Technology

Statistical Learning of the Giant Panda (Ailuropoda Melanoleuca) Ethology

Yixuan Liu

Instructor: Prof. Yifang Ma

2020-5-27

Contents

SUSTech

Background of the Giant Panda Research

Scientific facts

Diet, predators, conservation, ecology, human interactions, genes, biofuel and etc..

Living history

Evolution, population, diplomacy and etc..

Behavior research

Gender (Ding-zhen et al.), age (Hong et al.), vocalization (Charlton et al.), oestrus (Kleiman et al.) and etc..

Behavioral Research Methodology

SUSTech

- Exploratory statistics: correlation, scatter plots, and graphical visualization.
- Statistical inference and analysis:
  - parametric tests: t-test, z-test, ANOVA.
  - non-parametric tests: Mann-Whitney, chi-square test etc..
- Models for representing phenomenon: regression model, non-linear models, clustering, network models
- Newly promoted methods: random resampling, robust problems, missing data, meta-analysis, and other optimizations

Dataset Description

- From March 14^{th}, 2000 to July 28^{th}, 2000. 35 observation days, mostly 3 days between each interval. On observed days, use scanning method.

Internal Relationships

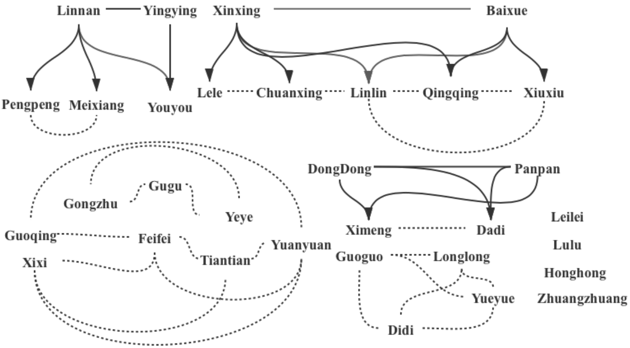


Figure: **Relationships among observed individuals.** Dashed lines for siblings, and solid lines for kinship.

Graphical displays

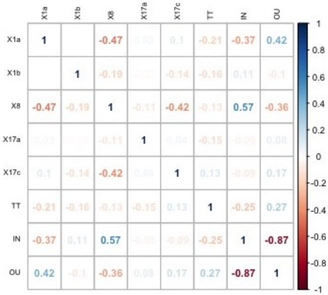


Figure: **Correlation plot.**
8 highest variables selected, only $\rho_{inout} = -0.87 < -0.8$.

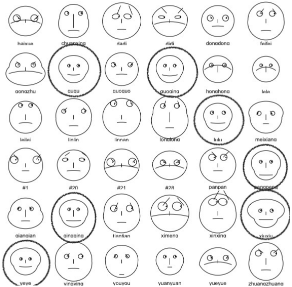


Figure: **Chernoff faces**

Autoregressive Integrated Moving Average (ARIMA)

- Use panda No.20's eating bamboos frequency (x_{1a}) with general class of model $ARIMA(p, d, q)$ including "autoregressive", "moving average" and "difference" terms are used for simulation.
- x_{1a} has autocorrelation function (ACF) almost truncated at $lag = 3$, partial autocorrelation function (PACF) decreased geometrically.
- $AIC = 151.72$,$AIC_C = 153.60$, $BIC = 159.91$ provides an $AR(3)$ model: $X_t = 0.18X_{t-1} + 0.15X_{t-2} + 0.44X_{t-3} + 2.88$
- Ljung-Box of forecasting residuals has Q-statistic equals 17.104 (p-value$= 0.646 > 0.05$). Residuals are white noise.

Autoregressive Integrated Moving Average (ARIMA)

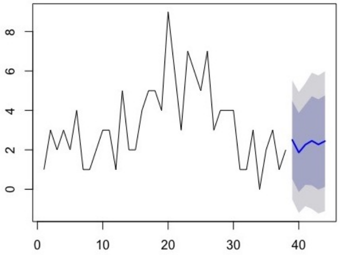


Figure: $AR(3)$ forecast of bamboo eating for panda No.20.

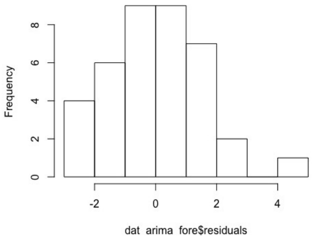


Figure: Residual distribution

Neural Network Auto-Regressive Model (NNAR)

- *ARIMA* forecast of rest performs poorly with $ARIMA(0,1,1)$ as exponential smoothing model.
- Procedures: lagged inputs are used in feed forward network, generating $z_j = b_j + \sum_{i=1}^{j} \omega_{i,j} x_i$ to the next hidden layer. Hidden layer uses $s(z) = \frac{1}{1+e^{-z}}$ as input to output layer, and reduce outliers. Output layer calculates back propagation errors to update $\omega_{i,j}$.
- The NNAR model retrieves results based on optimal number of lags according to AIC.
- Notation: NNAR(p, k) is a neural network with $\{y_{t-1}, y_{t-2}, \ldots, y_{t-p}\}$ as lagged inputs are k nodes in the hidden layer.

Prediction Intervals

- Predictions made through bootstrapped residuals.
- Fitted neural network: $y_t = f(y_{t-1}) + \epsilon_t$. Where f is a neural network 1 node in 1 hidden layer, the series $\{\epsilon_t\}$ are equal variance.
- Iteratively, by resampling ϵ_t from Gaussian distribution, $y^*_{T+1} = f(y_T) + \epsilon^*_{T+1}$,y^*_{T+2} =$f(y_{T+1}) + \epsilon^*_{T+2}, \cdots$. All possible future values are generated.

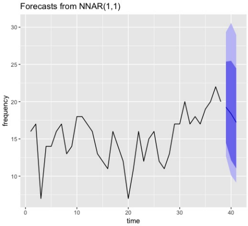


Figure: $NNAR(1,1)$ forecast of rest behavior

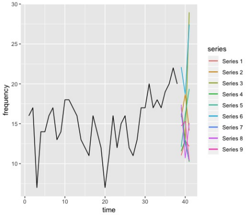


Figure: 9 future series

Bagging Time Series Model

- Bootstrapped time series: introduced uncertainty additionally from changing data generating model.
- For each bootstrapped series, an exponential smoothing $x_{i+1} = \alpha \sum_{j=0}^{i} (1-\alpha)^j x_{ij}$ is applied.
- $RMSE_{Bagging} = 6.14$, $RMSE_{NNAR} = 3.12$. $NNAR$ has better forcast. While prediction intervals of bagging forecast are always wider than others.

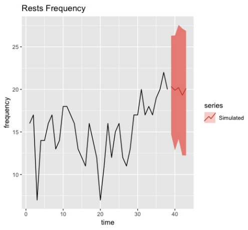


Figure: Bagging forecast

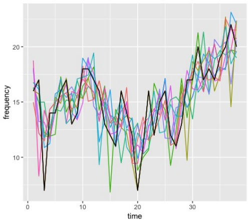


Figure: 10 bootstrapped series

Prediction Intervals

Table: Prediction band of **NNAR** forecast

Label	Forecast	Low 95	High 95
39	19.259	14.475	25.575
40	18.378	12.211	25.998
41	17.223	11.472	25.236

Table: Prediction band of **bagged** forecast

Label	Forecast	Low 95	High 95
39	20.087	14.036	26.443
40	19.882	13.813	25.450
41	20.086	12.607	26.772

- Bagged forecast has a wider prediction interval.

Sub-adults and Adults Individuals

Table: Effects of sex in captive sub-adults

Behavior	Male	Female	p-value
Eating bamboo (x_{1a})	4.25	3.55	0.2631
Rest (x_8)	9.75	12.90	0.0646
Investigating (x_{17a})	1.75	0.80	0.455

Table: Effects of sex in semi-ranging adults

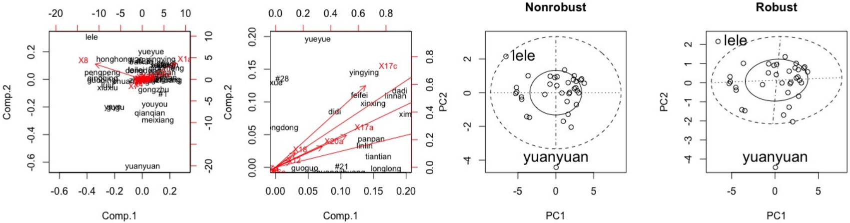
Behavior	Male	Female	p-value
Eating bamboo (x_{1a})	5.25	2.31	0.0211
Rest (x_8)	11.5	4.00	0.0002
Investigating (x_{17a})	0.75	0.365	0.5352

- Semi-ranging adults have significance difference in eating bamboos ($p\text{-}value = 0.0211$), and rest behavior ($p\text{-}value = 0.0002$).

Principal Components Analysis

Preparations for clustering

- The first 4 principal components account for 98.5% of variances (coefficients less than 0.4 omitted):
  - PC1: $0.551X_{1a} - 0.759X_8$ (eating bamboos versus rest)
  - PC2: $0.677X_{1a} + 0.635X_8$ (eating bamboos and rest)
  - PC3: $-0.771X_2 + 0.536X_{17c}$ (sitting versus walking)
  - PC4: $0.438X_{1a} - 0.407X_{1b} - 0.477X_2 - 0.511X_{17c}$ (eating bamboos versus eating others, walking and sitting)



SUSTech

Factor Analysis

Preparations for clustering

- Aim: find latent factors to simplify interpretation through oblique rotation from principal scores:

$$F_1^* = d_{11}F_1 + d_{12}F_2 + \ldots + d_{1m}F_m$$

$$F_2^* = d_{21}F_1 + d_{22}F_2 + \ldots + d_{2m}F_m$$

$$\ldots$$

$$F_m^* = d_{m1}F_1 + d_{m2}F_2 + \ldots + d_{mm}F_m$$

- Rough rule of thumb (Kaiser criterion) suggests 4 factors, accounting for 62% of total variance, with $\chi^2 = 41.17$, $p\text{-}value = 0.463 > 0.05$.

Output of Factor Scores

Preparations for clustering

Table: Loadings of 4 factors (< 0.5 omitted)

	Factor1	Factor2	Factor3	Factor4
x_{1a}	0.519	0.635	0.523	
x_{1b}	0.824			
x_2		0.795		
x_4	-0.684			
x_{6g1}		-0.560		
x_{6g2}			-0.734	
x_{6a}				-0.503
x_7	-0.801			
x_8				
x_{12}		0.605		
x_{17a}	0.662			
x_{17c}	0.945			
x_{18}		0.503		
x_{20a}		0.577		

SUSTech

Naming of Factors

Preparations for clustering

- Factor 1: **tense**. Eating bamboos, eating others, sitting straight and investigating versus climbing and playing. It shows restraint in movement and posture, carrying body stiffly.
- Factor 2: **oestrus**. Eating bamboos, pacing around, sniffing, drinking water and bleating versus licking. It is the intensity of demonstrated oestrus behavior.
- Factor 3: **oblivious**. Eating bamboos versus licking, unresponsive to events, and situations.
- Factor 4: **calm**. Minus scratching, not easily disturbed by changes in environment.

Agglomerative Clustering

- Intercluster dissimilarity measures:
  - single linkage: $d_{AB} = min\{d_{ij} | i \in A, j \in B\}$
  - complete linkage: $d_{AB} = max\{d_{ij} | i \in A, j \in B\}$
  - average linkage: $d_{AB} = n_A^{-1} n_B^{-1} \sum_{i \in A} \sum_{j \in B}$

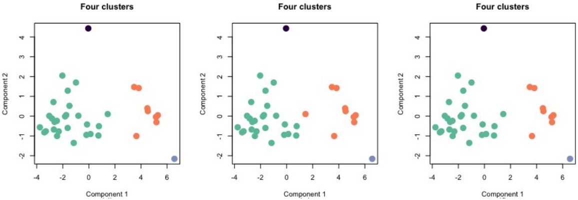


Figure: **Agglomerative clustering with 3 linkages** (from left to right: single, complete and average). Single linkage elongates, complete linkage creates ball-shaped, average linkage balance them two.

Partitional Clustering

- Methodology:
  - K-means
  - K-means++
  - Bisecting K-means
- Clustering evaluation criteria:
  - Average silhouette score (ASS): $s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$, where $a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \leq j} d(i,j)$, $b(i) = min_{k \leq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$.
  - Error sum of squares (SSE): $SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x_k})^2$. Which is the same calculation as within cluster sum of squares (WSS) in this research.
  - Calinski-Harabasz index (CHI): $s(k) = \frac{tr(B_k)}{tr(W_k)} \frac{m-k}{k-1}$, where m points have k clusters, B_k is between cluster covariance matrix, W_k as within-cluster covariance.

K-means Clustering

- Probability of a point belonging to each cluster: $\prod_{j=1}^{k}\prod_{i=1}^{N_j}\frac{1}{\sigma^2}exp(-\frac{||x_i-\mu_j||^2}{2\sigma^2})$
- Loss function: $J(\mu_1,\mu_2,\ldots,\mu_k)=\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_i-\mu_j)^2$.
- Cluster centroids: $\mu_j=\frac{\sum_{i=1}^{N_j}x_i}{N_j}$.
- Procedures:
  - (1): Select initial partition from agglomerative hierarchical clustering with average linkage.
  - (2): Calculates SSE (loss function) in each step.
  - (3): Repeat step (2) by yielding the largest improvement until no changes occur.

K-means++ Clustering

- K-means++ improves the initialization of clustering centers by careful seeding.
- Procedures:
  - (1): Randomly select a point from the dataset as the initial position c_i.
  - (2): Calculate smallest distance between the point to the closest center $D(x)$. Then, select the next center c_i with probability $\frac{D(x)^2}{\sum_x D(x)^2}$.
  - (3): continue with same process as (2) and (3) in k-means clustering.

SUSTech

Bisecting K-means Clustering

- Bisecting K-means is a hybrid algorithm between hierarchical clustering and K-means. It improves calculation efficiency by bisecting through k-means.
- Procedures:
  - (1): Compute the centroid w of the dataset, select a point c_L randomly and compute $c_R = w - (c_L - w)$.
  - (2): Divide the data M into two clusters M_L and M_R according to:

$$\begin{cases} x_i \in M_L, when \ ||x_i - c_i|| \leq ||x_i - c_R|| \\ x_i \in M_R, when \ ||x_i - c_i|| > ||x_i - c_R|| \end{cases} \tag{1}$$

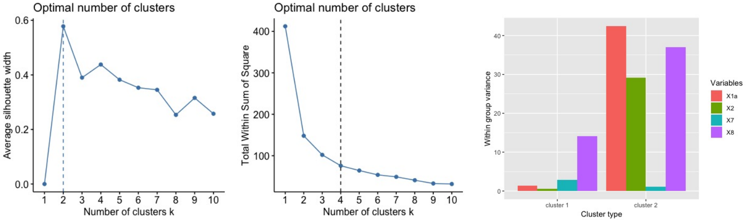
  - (3): Calculate centroids of M_L and M_R, noted as w_L and w_R.
  - (4): If $w_L = c_L$ and $w_R = c_R$ then stop, else repeat steps (2) and (3).

SUSTech

K-means Clustering Results

Best clustering of mean behavior

- $ASS = 0.58$, $SSE = 148.095$, $CHI = 60.68$ suggests a 2-means clustering.
- Cluster 1: labeled as **inactive**, for few variances of eating bamboos, $s_1(x_{1a}) = 1.36$ and walking around, $s_1(x_2) = 0.62$ comparing with rest, $s_1(x_8) = 14.09$.
- Cluster 2: labeled as **active**, large variance in both eating bamboo, walking around and having rest.



SUSTech

K-means Clustering Results

Best clustering of mean behavior

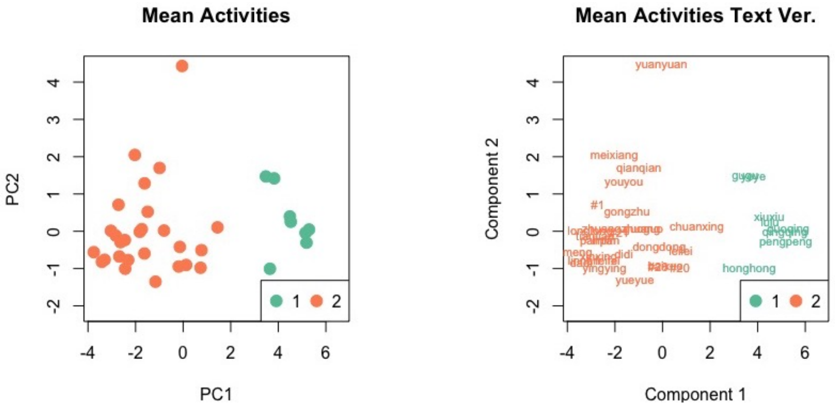


Figure: **Results of 2-means clustering.** Projected on two PCs.

SUSTech

DBSCAN

- *MinPts*: minimum number of points clustered together in a specific neighbor. When $MinPts \leq 2$, the result is same as agglomerative hierarchical clustering.
- EPS (ϵ): distance that contains *MinPts* neighbour points.
- Procedures:
  - (1) Randomly pick up a point c_L from the dataset.
  - (2) If at least *MinPts* points placed within neighbor with distance ϵ, labeled as the same cluster.
  - (3) Iteratively repeat process until every point picked.

HDBSCAN

- Improvements:
  - (1) Mutual reachability distance:
    $d_{mreach-k}(a,b) = max\{core_k(a), core_k(b), distance(a,b)\}$
  - (2) Minimum spanning tree (MST)
  - (3) Stability of cluster C_i: $\sum_{x_j \in C_i}(\lambda_{x_j} - \lambda_{birth})$
- Procedures:
  - (1) Compute $d_{mreach-k}$ for all points in the dataset.
  - (2) Compute the MST based on mutual reachability graph.
  - (3) Extend MST with edges, connecting to MST_{ext}.
  - (4) Make dendrogram and cut tree by extracting HDBSCAN hierarchy MST_{ext}.

HDBSCAN Results

A possible clustering of factor scores

- $CHI = 21.24$, $ASS = 0.27$.
- Top 5 members: Meixiang, Tiantian, Feifei. Linlin, Youyou.
- Top 5 outliers: Quoqing, Lele, No.1, Yueyue, Xinxing.

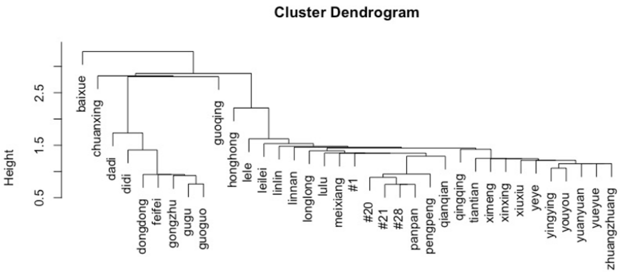


Figure: **HDBSCAN Dendrogram.** Two obvious clusters are fond.

HDBSCAN Results

A possible clustering of factor scores

- Cluster 1: labeled as **unhealthy**. Less tense and low oestrus behavior intensity.
- Cluster 2: labeled as **healthy**. High tense and above average oestrus behavior intensity. 72.2% of all individuals.

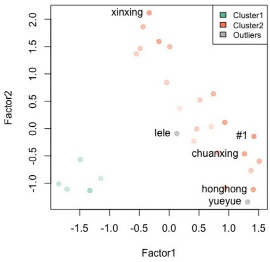


Figure: HDBSCAN of 2 clusters.

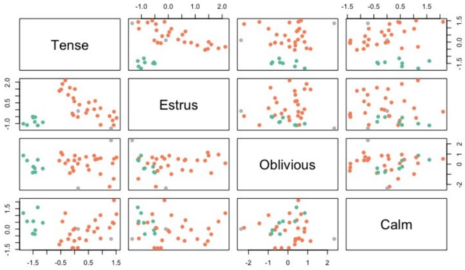


Figure: Scatterplot Matrix.

Spectral Clustering

- Developed from the graph theory, it identifies communities from the links between them.
- Procedures:
  - (1) Calculate the affinity matrix A by

$$\begin{cases} A_{ij} = exp(\frac{||s_i - s_j||^2}{2\sigma^2}), & when\ i \neq j \\ A_{ij} = 0, & when\ i = j \end{cases} \tag{2}$$

  - (2) Calculate the matrix $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D has sum of matrix A's.
  - (3) Find first k eigenvalues and eigenvectors, normalize to form matrix X.
  - (4) Apply k-means to each row of X for clustering.

Spectral Clustering Results

Best clustering of factor scores

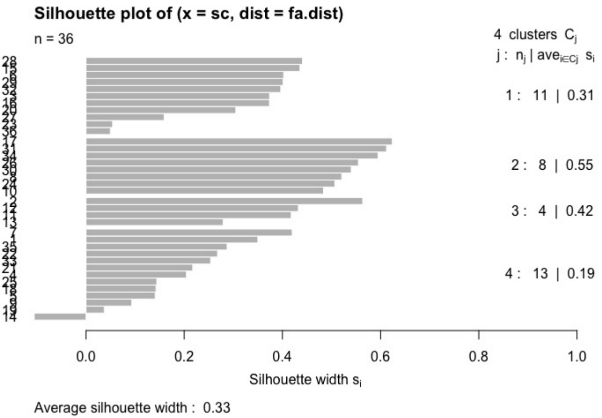


Figure: **Silhouette Plot.** With only abnormality of Linlin.

Spectral Clustering Results

Best clustering of factor scores

- Cluster 1: less tense, high intensity of oestrus behavior.
- Cluster 2: high tense and oestrus behavior. Mostly age in this group is larger than 6 (adults).
- Cluster 3: highly tense, less oestrus and oblivious.
- Cluster 4: highly tense, less oestrus and more oblivious. Age is significantly less than cluster 3 ($p\text{-}value = 0.034 < 0.05$).

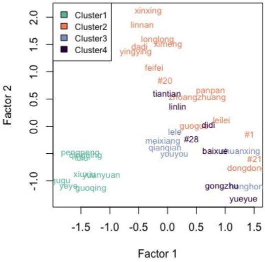


Figure: Spectral clustering results

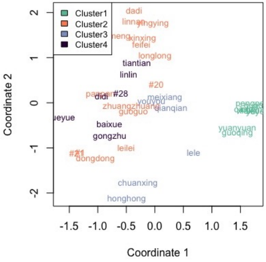


Figure: Results through MDS

Distance Measures

- Euclidean distance: $d(x, y) = ||x, y||_2$.
- Dynamic time warping: $DTW_p(x, y) = (\sum \frac{m_\phi lcm(k)^p}{M_\phi})^{\frac{1}{p}}$. Where lcm is the local cost matrix.
- Shape-based distance: $SBD(x, y) = 1 - \frac{max(NCC_c(x,y))}{||x||_2||y||_2}$. Where NCC_c is the cross-correlation with coefficient normalization of two time-series.
- Global alignment kernel: $k_{GA}(x, y) = \sum_\pi \prod_{i=1}^{|\pi|} \kappa(x_{\pi_1(i)}, y_{\pi_2(i)})$. Where κ is the local similarity function. Triangular global kernel is used to reduce GA kernel's complexity.

Intrinsic Measures of Clustering

For data not labeled in advance:

- Score function: $SF(c) = 1 - \frac{1}{e^{e^{betweeness+within}}}$
- Davies-Bouldin index: $DB(C) = \frac{1}{k}\sum max\{\frac{S(C_k)+S(C_l)}{d(\bar{C}_k,\bar{C}_l)}\}$. Where $S(C_k) = \frac{1}{|C_k|}\sum d(x_i, \bar{C}_k)$.
- Dunn index: $D(C) = \frac{min_{c_k\in c}\{min_{c_l\in C}\delta(c_k)\}}{max_{c_k\in c}\{\Delta(c_k)\}}$.
- COP index: $COP(C) = \frac{1}{N}\sum \frac{\sum d(x_i,\bar{C}_k)}{|C_k| min_{x_i\notin C_k} max_{x_j\notin C_k} d(x_i.x_j)}$.
- For fuzzy clustering using other criteria: MPC, K, T, SC and PBMF.

SUSTech

Clustering Outcomes

Table: intrinsic criteria of clustering

	ASS	SF	CHI	DB	D	COP
Hclus+L_2	0.12	0.00	10.67	1.61	0.64	0.64
Hclus+SBD	0.19	0.35	3.33	1.16	0.67	0.67
P+DTW	0.40	0.11	23.35	0.88	0.24	0.45
P+DTW_2+DBA	0.25	0.00	29.29	1.52	0.39	0.48
k-shape	0.11	0.41	17.96	1.63	0.39	0.55
P+GAK	0.60	0.63	49.64	0.34	0.11	0.18

- P: partitional clustering; Hclus: hierarchical clustering; L_2: Euclidean distance.

Partitional clustering using GAK Distance

Best time-series clustering of rest behavior

- Procedures:
  - (1) randomly select a series from the dataset as initial position c_i.
  - (2) calculate smallest distance between the series by GAK distance. Then update the cluster centroids.
  - (3) repeat step (2) until no improvement occurs.
- Cluster results:
  - Cluster 1: have fluctuated rest frequency. $\frac{7}{9}$ of the cluster are same to cluster 1 from spectral clustering.
  - Cluster 2: react to an upward variation of time. $\frac{26}{27}$ of the cluster are the same to cluster 2 from bisecting 2-means on mean behavior.

Partitional clustering using GAK Distance

Best time-series clustering of rest behavior

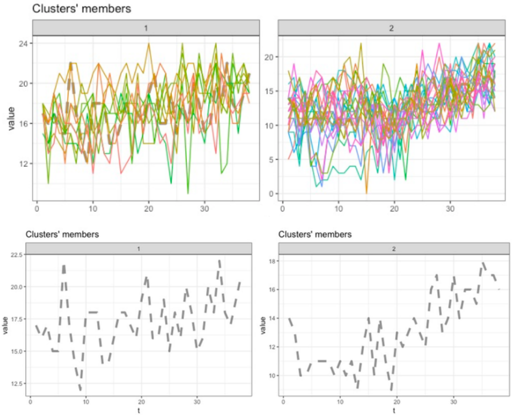


Figure: **Partitional clustering of GAK distance.** Grey lines are obtained prototypes.

Fuzzy Clustering

- Fuzzy clustering provides members of clusters to a certain degree. It is carried out through an iterative optimization of the objective function $\sum_{p=1}^{N}\sum_{c=1}^{k}\mu_{p,c}^{m}d_{p,c}^{2}$.
- Procedures:
  - (1) Initialize $U=[u_{p,c}]$ randomly.
  - (2) Calculate centers vectors $C_j=\frac{\sum_{p=1}^{N}\mu_{p,c}^{m}x_{p,i}}{\sum_{p=1}^{N}\mu_{p,c}^{m}}$.
  - (3) Update U by minimizing the objective function with GAK distance, until nearly no further improvements are made.
- Intrinsic measures $MPC=0.17$, $K=22.75$.

Fuzzy Clustering Outputs

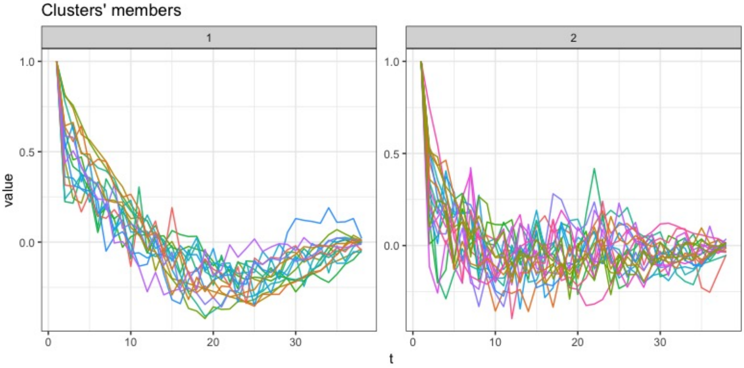


Figure: **Fuzzy clustering of GAK**. Lag-reaction to time is observed. Cluster 1: $\frac{2}{3}$ are female. Cluster 2: $\frac{11}{17}$ are male. Independent t-test has $p\text{-}value = 0.0167 < 0.05$.

Conclusion

- Work presented:
  - NNAR $\Rightarrow$ rest frequency prediction
  - Factor analysis $\Rightarrow$ 4 latent factors
  - K-means based on hierarchical clustering $\Rightarrow$ best clustering on mean behavior
  - Spectral clustering $\Rightarrow$ best clustering on factor scores
  - Partitional time series clustering with GAK $\Rightarrow$ rest variation with previous clustering result
  - Fuzzy time series clustering with GAK $\Rightarrow$ rest variation with gender
- Future research:
  - Integrate semi-supervision learning and sub-spaces of the giant pandas' behavior.
  - Naming of each cluster and factor.

Thank you for listening! Open for discussions.