A Cluster-Target Similarity Based Principal Component Analysis for Interval-Valued Data

University of Tsukuba
School of Systems and Information Engineering

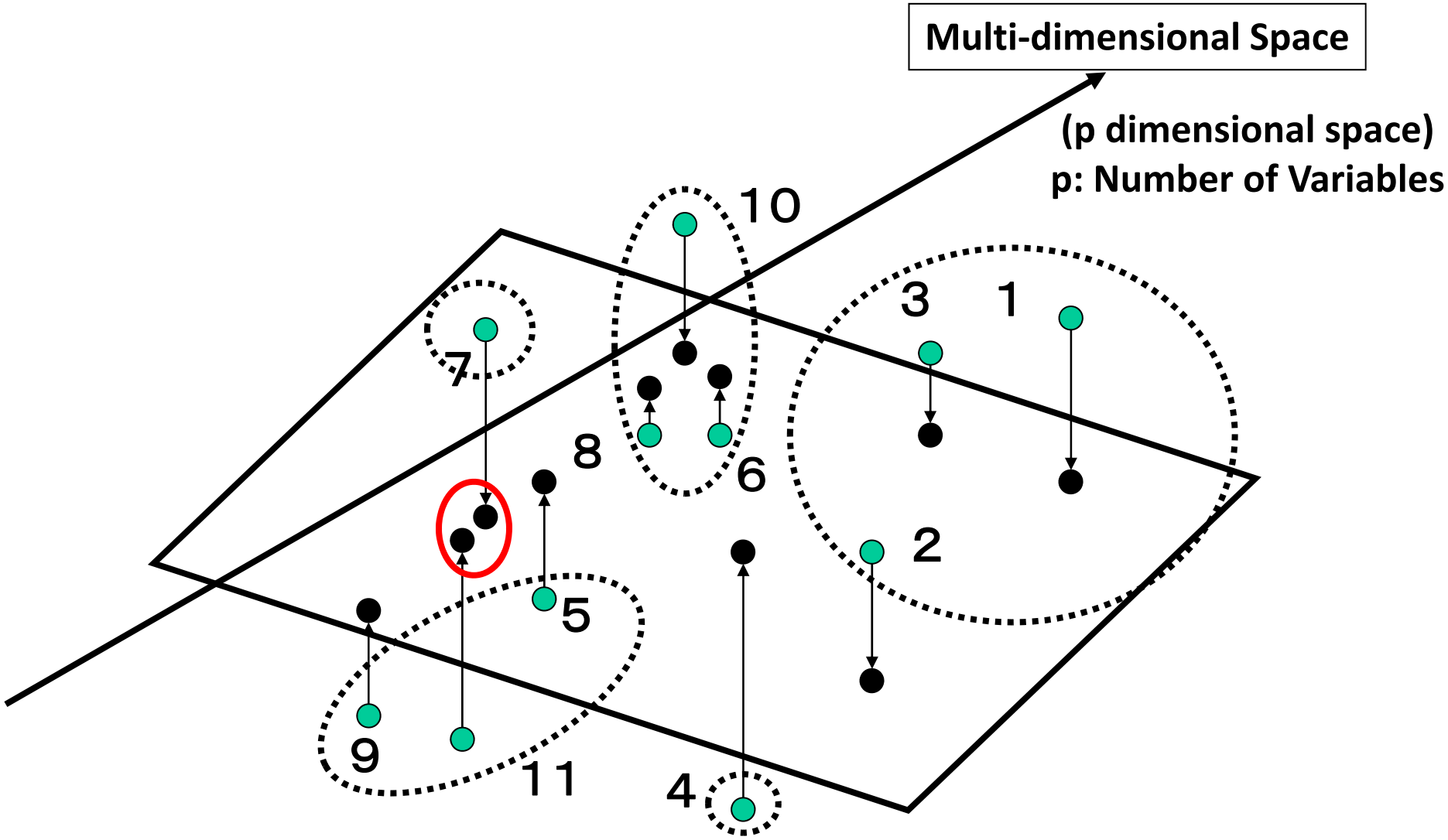
Mika Sato-Ilic

Energy Evaluation Data

Energy	JP1	JP2	…	UK1	UK2	…
1. Oil	[60,90]	[60,70]	…	[81,91]	[51,71]	…
2. Coal	[90,120]	[80,95]	…	[80,91]	[80,91]	…
3. Coal with CCS	[60,100]	[20,40]	…	[50,60]	[65,75]	…
4. Nuclear	[70,120]	[50,85]	…	[45,65]	[60,80]	…
5. Geothermal	[60,80]	[30,45]	…	[0,20]	[0,20]	…
6. Solar PV	[30,70]	[30,40]	…	[0,10]	[10,40]	…
7. Biomass	[40,100]	[20,35]	…	[60,70]	[60,70]	…
8. On. Wind, large	[70,100]	[50,60]	…	[60,72]	[50,70]	…
9. Mun/Ind Waste	[83,111]	[50,65]	…	[60,80]	[80,90]	…
10. Hydro	[70,100]	[40,60]	…	[65,75]	[60,70]	…
11. Gas	[80,120]	[65,85]	…	[87,97]	[87,97]	…

Joint Research: ESRC-funded Sussex Energy Group at SPRU (Science and Technology Policy Research, University of Sussex)
Sustainable Energy/Environment & Public Policy (SEPP), University of Tokyo
(ESRC: Economic and Social Research Council)

Principal Component Analysis based on Classification Structure by Fuzzy Clustering



Adaptable Classification Structure ⟺ Adaptable Number of Clusters

Principal Component Analysis for Metric Projection

Metric Projection

$$P_L : X \to L$$

$$P_L(\mathbf{x}) = \{\mathbf{y} \in L : \|\mathbf{x} - \mathbf{y}\| = d(\mathbf{x}, L)\}$$

$$\mathbf{x} \in X, \quad d(\mathbf{x}, L) = \inf_{\mathbf{y} \in L} \|\mathbf{x} - \mathbf{y}\|$$

X : **Inner Product Space**

L : **Nonempty Subset of X**

Principal Component Analysis for Metric Projection

Metric Projection

$$P_L : X \to L$$

P_L is nonexpansive

$$\left\| P_L(\mathbf{x}) - P_L(\mathbf{y}) \right\| \le \left\| \mathbf{x} - \mathbf{y} \right\|$$

$$\mathbf{x}, \mathbf{y} \in X$$

L : Convex Chebyshev Set

For each $\mathbf{x} \in X$, there exists at least one nearest point in L

Principal Component Analysis

$$C(\mathbf{x}, \mathbf{y}) = \left\| \mathbf{x} - \mathbf{y} \right\| - \left\| P_L(\mathbf{x}) - P_L(\mathbf{y}) \right\|$$

$\left\| \mathbf{x} - \mathbf{y} \right\|$: Dissimilarity of Objects

$\left\| P_L(\mathbf{x}) - P_L(\mathbf{y}) \right\|$: Dissimilarity of Objects on Projected Space

Principal Component Analysis

$$X = \begin{pmatrix} \tilde{\mathbf{x}}_1 \\ \vdots \\ \tilde{\mathbf{x}}_n \end{pmatrix}, \quad \tilde{\mathbf{x}}_i = (x_{i1}, \cdots, x_{ip}), \quad i = 1, \cdots, n$$

$$X = (\mathbf{x}_1, \cdots, \mathbf{x}_p), \quad \mathbf{x}_a = \begin{pmatrix} x_{1a} \\ \vdots \\ x_{na} \end{pmatrix}, \quad a = 1, \cdots, p$$

n : Number of Objects p : Number of Variables

Minimize $F = \sum_{a=1}^{p} (\mathbf{x}_a - X\mathbf{l}_1)'(\mathbf{x}_a - X\mathbf{l}_1)$

$\mathbf{l}_1 = (X'X)^{-1} X'\mathbf{x}_a$

$F^* = \sum_{a=1}^{p} (\mathbf{x}_a - X(X'X)^{-1}X'\mathbf{x}_a)'(\mathbf{x}_a - X(X'X)^{-1}X'\mathbf{x}_a)$

$X(X'X)^{-1}X'$: Projection to Subspace Spanned by $\mathbf{x}_1, \cdots, \mathbf{x}_p$

$$\sum_{a \neq b=1}^{p} \left((\mathbf{x}_a - \mathbf{x}_b) - X(X'X)^{-1}X'(\mathbf{x}_a - \mathbf{x}_b)\right)'\left((\mathbf{x}_a - \mathbf{x}_b) - X(X'X)^{-1}X'(\mathbf{x}_a - \mathbf{x}_b)\right)$$

Principal Component Analysis

$X(X'X)^{-1}X'$: Projection to Subspace Spanned by $\mathbf{x}_1, \cdots, \mathbf{x}_p$

$$P_X \equiv X(X'X)^{-1}X'$$

$P_X = P_X P_X$: *Idempotent* $\qquad$ $P_X' = P_X$: *Symmetry*

$$\mathbf{x}_a, \mathbf{x}_b \in V^n, \quad \exists \mathbf{x}_a - \mathbf{x}_b \in V^n$$

$$F^* = \sum_{a=1}^{p} (\mathbf{x}_a - P_X \mathbf{x}_a)'(\mathbf{x}_a - P_X \mathbf{x}_a) = \sum_{a=1}^{p} (\mathbf{x}_a' \mathbf{x}_a - \mathbf{x}_a' P_X \mathbf{x}_a)$$

$$\sum_{a \neq b=1}^{p} \left((\mathbf{x}_a - \mathbf{x}_b) - P_X(\mathbf{x}_a - \mathbf{x}_b)\right)' \left((\mathbf{x}_a - \mathbf{x}_b) - P_X(\mathbf{x}_a - \mathbf{x}_b)\right)$$

$$= \sum_{a \neq b=1}^{p} (\mathbf{x}_a - \mathbf{x}_b)'(\mathbf{x}_a - \mathbf{x}_b) - (\mathbf{x}_a - \mathbf{x}_b)' P_X (\mathbf{x}_a - \mathbf{x}_b)$$

$$= 2p \sum_{a=1}^{p} (\mathbf{x}_a' \mathbf{x}_a - \mathbf{x}_a' P_X \mathbf{x}_a) - 2 \sum_{a \neq b=1}^{p} (\mathbf{x}_a' \mathbf{x}_b - \mathbf{x}_a' P_X \mathbf{x}_b)$$

F^*

Covariance between Variables

Fuzzy Cluster based Principal Component Analysis

$$\sum_{a\neq b=1}^{p}\left((\mathbf{x}_a-\mathbf{x}_b)-P_X(\mathbf{x}_a-\mathbf{x}_b)\right)'\left((\mathbf{x}_a-\mathbf{x}_b)-P_X(\mathbf{x}_a-\mathbf{x}_b)\right)$$

$$=\sum_{a\neq b=1}^{p}(\mathbf{x}_a-\mathbf{x}_b)'(\mathbf{x}_a-\mathbf{x}_b)-(\mathbf{x}_a-\mathbf{x}_b)'P_X(\mathbf{x}_a-\mathbf{x}_b)$$

$$=2p\sum_{a=1}^{p}(\mathbf{x}_a'\mathbf{x}_a-\mathbf{x}_a'P_X\mathbf{x}_a)-2\sum_{a\neq b=1}^{p}(\mathbf{x}_a'\mathbf{x}_b-\mathbf{x}_a'P_X\mathbf{x}_b)$$

F^*

Covariance between Variables

Adaptable Classification Structure based on an Appropriate Number of Clusters

Dissimilarity Structure of Objects in Higher Dimensional Space

Selection of an Appropriate Number of Clusters

$$S \rightarrow$$

K : Number of Clusters

$$K = a \rightarrow U^{(a)} \rightarrow \tilde{S}^{(a)}$$
$$K = b \rightarrow U^{(b)} \rightarrow \tilde{S}^{(b)}$$
$$(a \neq b)$$

S : Observed Similarity Data

$U^{(l)}$: Classification Structure for S when the Number of Clusters is l

(A Result of Fuzzy Clustering) $l = 1, \cdots, K$

$\tilde{S}^{(l)}$: Restored Similarity by Using $U^{(l)}$

Select Closest $\tilde{S}^{(l)}$ to S among $\{\tilde{S}^{(2)}, \cdots, \tilde{S}^{(K)}\}$

Select Most Explainable Classification Structure for Original Data

Appropriate Number of Clusters is l

Asymmetric Similarity of Interval-Valued Data

$$Y = (y_{ia}) = ([\underline{y}_{ia}, \bar{y}_{ia}]), \quad i = 1, \cdots, n, \quad a = 1, \cdots, p$$

Dissimilarity between $\mathbf{y}_i = (y_{i1}, \cdots, y_{ip})$ *and* $\mathbf{y}_j = (y_{j1}, \cdots, y_{jp})$

$$d_{ij} = \sum_{a=1}^{p} \sup\{d(x, y_{ja}) \mid x \in y_{ia}\}, \quad d(x, y_{ja}) = \inf\{d(x, y) \mid y \in y_{ja}\}$$

$$d_{ji} = \sum_{a=1}^{p} \sup\{d(y_{ia}, y) \mid y \in y_{ja}\}, \quad d(y_{ia}, y) = \inf\{d(x, y) \mid x \in y_{ia}\}$$

$$d_{ij} \neq d_{ji} \quad (i \neq j)$$

$$s_{ij} = 1 - d_{ij} / \max_{i,j}\{d_{ij}\}, \quad i, j = 1, \cdots, n$$

$$s_{ij} \neq s_{ji} \quad (i \neq j)$$

Asymmetric Fuzzy Clustering Model

(Sato and Sato, 1995)

Asymmetric Similarity Data

$$S = (s_{ij}), \quad s_{ij} \neq s_{ji}, \quad (i \neq j), \quad i, j = 1,..., n$$

$$s_{ij} = \sum_{k=1}^{K} \sum_{l=1}^{K} w_{kl} u_{ik} u_{jl} + \varepsilon_{ij}, \quad i, j = 1,..., n$$

s_{ij} : *Asymmetric Similarity Between Objects i and j*

u_{ik} : *Degree of Belongingness of an Ojbect i to a Cluster k*

w_{kl} : *Asymmetric Similarity Between Clusters k and l*

ε_{ij} : *Error*

$w_{kl} \neq w_{lk}$ $\quad s_{ij} \neq s_{ji}$ $\quad u_{ik} \in [0,1], \quad \sum_{k=1}^{K} u_{ik} = 1, \quad m \in (1, \infty)$

n : *Number of Objects* $\quad K$: *Number of Clusters*

Asymmetric Similarity of Clusters

$$w_{kl}^{(K)} = 1 - \frac{1}{1 + \exp^{-\tilde{w}_{kl}^{(K)}}}, \quad k, l = 1, \ldots, K$$

$$\tilde{w}_{kl}^{(K)} = \frac{1}{2}\left(\left(\left\| \boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)} \right\|_{\Sigma_{(k,K)}^{-1}} \right) + tr\left(\Sigma_{(k,K)}^{-1} \Sigma_{(l,K)} - I \right) + \log\left(\frac{|\Sigma_{(k,K)}|}{|\Sigma_{(l,K)}|} \right) \right)$$

$$\left\| \boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)} \right\|_{\Sigma_{(k,K)}^{-1}} = (\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)})' \Sigma_{(k,K)}^{-1} (\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)})$$

$\boldsymbol{\mu}_{(k,K)}$: Expected Value of Data in Cluster k

$\Sigma_{(k,K)}$: Variance – Covariance Matrix for Cluster k

$\tilde{w}_{kl}^{(K)} \neq \tilde{w}_{lk}^{(K)}, \quad (k \neq l), \quad \tilde{w}_{kl}^{(K)} \in [0,1]$

K : Number of Clusters

Criterion for Selection of Number of Clusters

$$C(K) = \frac{\sum_{i \neq j=1}^{n} s_{ij} \tilde{s}_{ij}^{(K)}}{\sqrt{\sum_{i \neq j=1}^{n} s_{ij}^2} \sqrt{\sum_{i \neq j=1}^{n} \tilde{s}_{ij}^{(K)^2}}}$$

Alignment **Degree of Agreement**

$$\tilde{s}_{ij}^{(K)} = \sum_{k=1}^{K} \sum_{l=1}^{K} w_{kl}^{(K)} u_{ik}^{(K)} u_{jl}^{(K)}, \quad i, j = 1,..., n$$

Restored Asymmetric Similarity

K : *Number of Clusters* $\quad k = 1,..., K$

Concentration around Expected Value for the Different w_{kl}

$$C(K) = \frac{\sum_{i \neq j=1}^{n} s_{ij} \tilde{s}_{ij}^{(K)}}{\sqrt{\sum_{i \neq j=1}^{n} s_{ij}^2} \sqrt{\sum_{i \neq j=1}^{n} \tilde{s}_{ij}^{(K)^2}}}$$

$$P(|\hat{C}(s, \hat{s}, w) - E[\hat{C}(s, \hat{s}, w)]| \geq \varepsilon) \leq 2 \exp\left(\frac{-m^2 a^4 \varepsilon^2}{c} \right)$$

$$\hat{C}(s_{ij}, \hat{s}_{ij}, \hat{w}_{kl}) = \hat{C}(s, \hat{s}, \hat{w})$$

Concentration around Expected Value for the Different w_{kl}

$$|\hat{C}(s,\hat{s},\hat{w})-\hat{C}(s,\tilde{s},\tilde{w})|\le\frac{6}{ma^2},\ \ m=n(n-1)$$

$$\hat{C}(s,\tilde{s},\tilde{w})=\hat{C}(s,\tilde{s}+d\tilde{s},\hat{w})$$

$$\hat{C}(s,\tilde{s}+d\tilde{s},\hat{w})-\hat{C}(s,\tilde{s},\hat{w})\le\frac{2\langle\tilde{s},d\tilde{s}\rangle}{\|\tilde{s}\|(\|\tilde{s}\|+\|\tilde{s}+d\tilde{s}\|)}+\frac{\|d\tilde{s}\|^2}{\|\tilde{s}\|(\|\tilde{s}\|+\|\tilde{s}+d\tilde{s}\|)}+\left\langle\frac{d\tilde{s}}{\|\tilde{s}\|},\frac{s}{\|s\|}\right\rangle$$

$$\langle d\tilde{s},\tilde{s}\rangle\le\frac{2}{m},\ \ m=n(n-1)$$

$$\hat{C}(s_{ij},\hat{s}_{ij},\hat{w}_{kl})=\hat{C}(s,\hat{s},\hat{w}),\quad \hat{C}(s_{ij},\tilde{s}_{ij},\tilde{w}_{kl})=\hat{C}(s,\tilde{s},\tilde{w})$$

$$K:\ Number\ of\ Clusters\qquad k=1,\ldots,K$$

Theorem: (C. McDiarmid)

$$\sup_{x_1,\cdots,x_n,\hat{x}_i} |\, f(x_1,\cdots,x_n) - f(x_1,\cdots,x_{i-1},\hat{x}_i,x_{i+1},\cdots,x_n) \,| \le c_i$$

$$P(|\, f(X_1,\cdots,X_n) - Ef(X_1,\cdots,X_n) \,| \ge \varepsilon) \le 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right), \ \textit{for all } \varepsilon > 0$$

$X_1,\cdots,X_n$: *Independent Random Variables Taking Values in a Set* A

$f : A^n \to R$ *satisfies for* $1 \le i \le n$

$$|\, \hat{C}(s,\hat{s},\hat{w}) - \hat{C}(s,\tilde{s},\tilde{w}) \,| \le \frac{6}{ma^2}, \ \ m = n(n-1)$$

$$P(|\, \hat{C}(s,\hat{s},w) - E[\hat{C}(s,\hat{s},w)] \,| \ge \varepsilon) \le 2\exp\left(\frac{-m^2a^4\varepsilon^2}{c}\right)$$

Fuzzy Cluster based Principal Component Analysis

$$\sum_{a\neq b=1}^{p}\left((\mathbf{x}_a-\mathbf{x}_b)-P_X(\mathbf{x}_a-\mathbf{x}_b)\right)'\left((\mathbf{x}_a-\mathbf{x}_b)-P_X(\mathbf{x}_a-\mathbf{x}_b)\right)$$

$$=\sum_{a\neq b=1}^{p}(\mathbf{x}_a-\mathbf{x}_b)'(\mathbf{x}_a-\mathbf{x}_b)-(\mathbf{x}_a-\mathbf{x}_b)'P_X(\mathbf{x}_a-\mathbf{x}_b)$$

$$=2p\sum_{a=1}^{p}(\mathbf{x}_a'\mathbf{x}_a-\mathbf{x}_a'P_X\mathbf{x}_a)-2\sum_{a\neq b=1}^{p}(\mathbf{x}_a'\mathbf{x}_b-\mathbf{x}_a'P_X\mathbf{x}_b)$$

F^*

Covariance between Variables

Adaptable Classification Structure based on an Appropriate Number of Clusters

Dissimilarity Structure of Objects in Higher Dimensional Space

Fuzzy Cluster based Covariance Matrix with respect to Variables

$$C = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})^t(\mathbf{x}_i - \bar{\mathbf{x}})}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m}}, \quad \bar{\mathbf{x}} = \frac{\sum_{i=1}^{n}\mathbf{x}_i}{n}$$

$$u_{ik} \in [0,1], \quad \sum_{k=1}^{K} u_{ik} = 1, \quad m \in (1,\infty)$$

Fuzzy Clustering

Special case

$$C = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^t(\mathbf{x}_i - \bar{\mathbf{x}})}{n}$$

$$u_{ik} \in \{0,1\}, \quad \sum_{k=1}^{K} u_{ik} = 1$$

Hard Clustering

***n* : Number of Objects** ***K* : Number of Clusters**

Fuzzy Cluster based Covariance Matrix with respect to Variables

$$C = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})^t(\mathbf{x}_i - \bar{\mathbf{x}})}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m}}, \quad \bar{\mathbf{x}} = \frac{\sum_{i=1}^{n}\mathbf{x}_i}{n}$$

$$u_{ik} \in [0,1], \quad \sum_{k=1}^{K} u_{ik} = 1, \quad m \in (1,\infty)$$

$$C = (c_{ab}), \quad c_{ab} = \sum_{i=1}^{n} w_i (x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b), \quad a,b = 1,\cdots,p$$

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_{ia}}{n}, \qquad w_i = \frac{\sum_{k=1}^{K} u_{ik}^{m}}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m}}$$

n **: Number of Objects** ***K*** **: Number of Clusters**

Fuzzy Cluster based Covariance Matrix with respect to Variables

$$C = (c_{ab}), \quad c_{ab} = \sum_{i=1}^{n} w_i (x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b), \quad a, b = 1, \cdots, p$$

$$w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^m}$$

$$u_{ik} \in [0,1], \quad \sum_{k=1}^{K} u_{ik} = 1 \qquad m \in (1, \infty)$$

Crisp Classification of an Object i → w_i Becoms Larger

Uncertainty Classification of an Object i

→ w_i Becomes Smaller

n : Number of Objects K : Number of Clusters

Fuzzy Cluster based Covariance Matrix with respect to Variables

$$C = (c_{ab}), \quad c_{ab} = \sum_{i=1}^{n} w_i (x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b), \quad a, b = 1, \cdots, p$$

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_{ia}}{n}, \quad w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m} \qquad m \in (1, \infty)$$

Fuzzy Clustering

$$u_{ik} \in [0,1], \quad \sum_{k=1}^{K} u_{ik} = 1 \quad \Rightarrow \quad w_i > 0, \quad \sum_{i=1}^{n} w_i = 1$$

Hard Clustering

$$u_{ik} \in \{0,1\}, \quad \sum_{k=1}^{K} u_{ik} = 1 \quad \Rightarrow \quad w_i = \frac{1}{n}, \quad \forall i$$

n : Number of Objects K : Number of Clusters

Fuzzy Cluster based Covariance Matrix for Interval-Valued Data

$$C = (c_{ab}), \quad c_{ab} = \sum_{i=1}^{n} w_i (x_{ia} - \overline{x}_a)(x_{ib} - \overline{x}_b), \quad a, b = 1, \cdots, p$$

$$\overline{x}_a = \frac{\sum_{i=1}^{n} x_{ia}}{n}, \quad w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^m}$$

n **: Number of Objects** ***K*** **: Number of Clusters**

Fuzzy Cluster based Covariance When x_{ia} are Interval-Valued Data ?

Empirical Joint Function for Interval-Valued Data

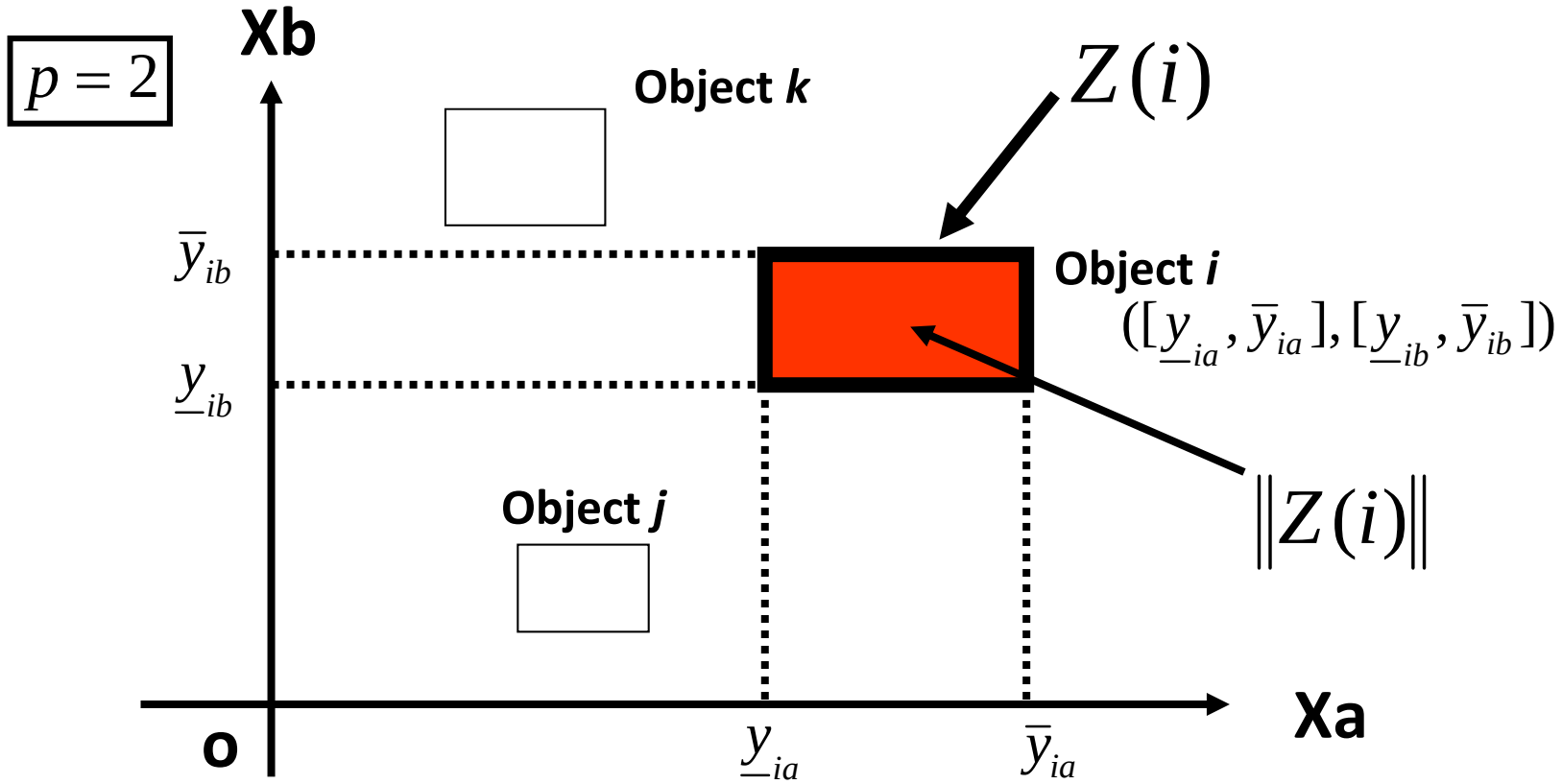
(Bertrand and Goupil, 2000)

$$f(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n} \frac{I_i(y_a, y_b)}{\|Z(i)\|},$$

$$I_i(y_a, y_b) = \begin{cases} 1, & ([\underline{y}_{ia}, \overline{y}_{ia}], [\underline{y}_{ib}, \overline{y}_{ib}]) \in Z(i) \\ 0, & \textit{Otherwise} \end{cases}, \quad \forall \underline{y}_{ia}, \overline{y}_{ia}, \underline{y}_{ib}, \overline{y}_{ib}$$

Uniform Distribution for Each Interval $[\underline{y}_{ia}, \overline{y}_{ia}]$

$p = 2$



Fuzzy Cluster based Covariance Matrix for Interval-Valued Data

$$C = (c_{ab}), \quad c_{ab} = \sum_{i=1}^{n} w_i (x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b), \quad a, b = 1, \cdots, p$$

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_{ia}}{n}, \quad w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m}$$

When *xia* are Interval-Valued Data ?

$$\hat{C} = (\hat{c}_{ab}), \quad \hat{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a - \bar{y}_a)(y_b - \bar{y}_b)\, \tilde{f}(y_a, y_b)\, dy_a dy_b$$

$$\bar{y}_a = \frac{1}{2n}\sum_{i=1}^{n} (\underline{y}_{ia} + \bar{y}_{ia}) : \text{Symbolic Empirical Mean}$$

$$\tilde{f}(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n} \frac{w_i I_i(y_a, y_b)}{\|Z(i)\|} : \text{Weighted Empirical Joint Function}$$

Fuzzy Cluster based Covariance Matrix for Interval-Valued Data

$$\hat{C} = (\hat{c}_{ab})$$

$$\hat{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y_a - \bar{y}_a)(y_b - \bar{y}_b)\,\tilde{f}(y_a, y_b)\,dy_a\,dy_b$$

$$= \frac{1}{4n}\sum_{i=1}^{n} w_i(\bar{y}_{ia} + \underline{y}_{ia})(\bar{y}_{ib} + \underline{y}_{ib}) - \frac{1}{n}\bar{y}_b\sum_{i=1}^{n}\frac{w_i(\bar{y}_{ia} + \underline{y}_{ia})}{2} - \frac{1}{n}\bar{y}_a\sum_{i=1}^{n}\frac{w_i(\bar{y}_{ib} + \underline{y}_{ib})}{2} + \frac{1}{n}\bar{y}_a\bar{y}_b$$

$$w_i = \frac{\sum_{k=1}^{K} u_{ik}^{m}}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m}}, \quad m \in (1, \infty)$$

PCA based on Fuzzy Covariance for Interval-Valued Data

$$\mathbf{l}_1' = (l_{11}, l_{12}, \cdots, l_{1p})$$

Corresponding Eigen-Vector For the Maximum Eigen-Value of $\hat{C}$

$\mathbf{z}_1 = \bar{Y}\mathbf{l}_1$: **First Principal Component**

$$\bar{Y} = (\bar{y}_{ia}), \quad \bar{y}_{ia} = \frac{\underline{y}_{ia} + \bar{y}_{ia}}{2}$$

$$i = 1, \cdots, n, \quad a = 1, \cdots, p$$

Conventional PCA for Interval-Valued Data

Centers Method

$$\tilde{Y} = \begin{pmatrix} y_{11}^c & \cdots & y_{1p}^c \\ \vdots & \ddots & \vdots \\ y_{n1}^c & \cdots & y_{np}^c \end{pmatrix}, \quad y_{ia}^c = \frac{\underline{y}_{ia} + \overline{y}_{ia}}{2}, \quad y_{ia} = [\underline{y}_{ia}, \overline{y}_{ia}]$$

$$i = 1, \cdots, n, \;\; a = 1, \cdots p$$

$$\boxed{\mathrm{cov}\{\tilde{Y}\} = \tilde{C}}$$

Covariance Matrix for Interval-Valued Data

(Billard and Diday, 2000)

$$\tilde{C} = (\tilde{c}_{ab}), \quad \tilde{c}_{ab} = \frac{1}{4n} \sum_{i=1}^{n} (\underline{y}_{ia} + \overline{y}_{ia})(\underline{y}_{ib} + \overline{y}_{ib}) - \overline{y}_a \overline{y}_b$$

$$\boxed{f(y_a, y_b) = \frac{1}{n} \sum_{i=1}^{n} \frac{I_i(y_a, y_b)}{\|Z(i)\|}}$$ (Bertrand and Goupil, 2000)

Principal Component Analysis : Centers Method

Comparison between Proposed PCA and Conventional PCA for Interval-Valued Data

$$\hat{C} = (\hat{c}_{ab}), \quad \hat{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a - \bar{y}_a)(y_b - \bar{y}_b)\, \tilde{f}(y_a, y_b)\, dy_a dy_b$$

$$\tilde{f}(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n} \frac{w_i I_i(y_a, y_b)}{\|Z(i)\|} : \text{Weighted Empirical Joint Function}$$

$$w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m}, \quad m \in (1, \infty)$$

Fuzzy Clustering

$$\tilde{C} = (\tilde{c}_{ab}), \quad \tilde{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a - \bar{y}_a)(y_b - \bar{y}_b)\, f(y_a, y_b)\, dy_a dy_b$$

$$f(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n} \frac{I_i(y_a, y_b)}{\|Z(i)\|} : \text{Empirical Joint Function}$$

Centers Method

$$w_i = 1, \quad \forall i$$

Hard Clustering

Fuzzy Cluster based Covariance Matrix with respect to Variables

$$C = (c_{ab}), \quad c_{ab} = \sum_{i=1}^{n} w_i (x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b), \quad a, b = 1, \cdots, p$$

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_{ia}}{n}, \quad w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m} \qquad m \in (1, \infty)$$

$$u_{ik} \in [0,1], \quad \sum_{k=1}^{K} u_{ik} = 1 \quad \Rightarrow \quad w_i > 0, \quad \sum_{i=1}^{n} w_i = 1$$

$$u_{ik} \in \{0,1\}, \quad \sum_{k=1}^{K} u_{ik} = 1 \quad \Rightarrow \quad w_i = \frac{1}{n}, \quad \forall i$$

n **: Number of Objects** ***K*** **: Number of Clusters**

Energy Evaluation Data

Energy	JP1	JP2	…	UK1	UK2	…
1. Oil	[60,90]	[60,70]	…	[81,91]	[51,71]	…
2. Coal	[90,120]	[80,95]	…	[80,91]	[80,91]	…
3. Coal with CCS	[60,100]	[20,40]	…	[50,60]	[65,75]	…
4. Nuclear	[70,120]	[50,85]	…	[45,65]	[60,80]	…
5. Geothermal	[60,80]	[30,45]	…	[0,20]	[0,20]	…
6. Solar PV	[30,70]	[30,40]	…	[0,10]	[10,40]	…
7. Biomass	[40,100]	[20,35]	…	[60,70]	[60,70]	…
8. On. Wind, large	[70,100]	[50,60]	…	[60,72]	[50,70]	…
9. Mun/Ind Waste	[83,111]	[50,65]	…	[60,80]	[80,90]	…
10. Hydro	[70,100]	[40,60]	…	[65,75]	[60,70]	…
11. Gas	[80,120]	[65,85]	…	[87,97]	[87,97]	…

Joint Research: ESRC-funded Sussex Energy Group at SPRU (Science and Technology Policy Research, University of Sussex)
Sustainable Energy/Environment & Public Policy (SEPP), University of Tokyo
(ESRC: Economic and Social Research Council)

Asymmetric Dissimilarity

Objects	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.70	0.78	0.83	0.38	0.36	0.66	0.78	0.78	0.70	0.77
2	0.67	1.00	0.51	0.81	0.05	0.03	0.37	0.50	0.63	0.43	0.80
3	0.73	0.49	1.00	0.77	0.45	0.50	0.77	0.73	0.75	0.70	0.57
4	0.70	0.70	0.70	1.00	0.22	0.22	0.51	0.62	0.67	0.54	0.67
5	0.44	0.14	0.58	0.42	1.00	0.78	0.64	0.59	0.46	0.62	0.26
6	0.29	0.00	0.49	0.30	0.64	1.00	0.43	0.36	0.30	0.41	0.12
7	0.67	0.39	0.81	0.63	0.63	0.50	1.00	0.82	0.66	0.81	0.51
8	0.80	0.55	0.81	0.78	0.55	0.46	0.85	1.00	0.80	0.81	0.66
9	0.80	0.69	0.84	0.82	0.42	0.40	0.70	0.81	1.00	0.79	0.76
10	0.76	0.53	0.84	0.75	0.62	0.55	0.86	0.86	0.83	1.00	0.64
11	0.75	0.82	0.61	0.80	0.19	0.17	0.50	0.64	0.73	0.56	1.00

1:	Oil	3:	Coal with CCS	5:	Geothermal	7:	Biomass	9:	Mun/Ind Waste	11: Gas
2:	Coal	4:	Nuclear	6:	Solar PV	8:	On. Wind, large	10:	Hydro	

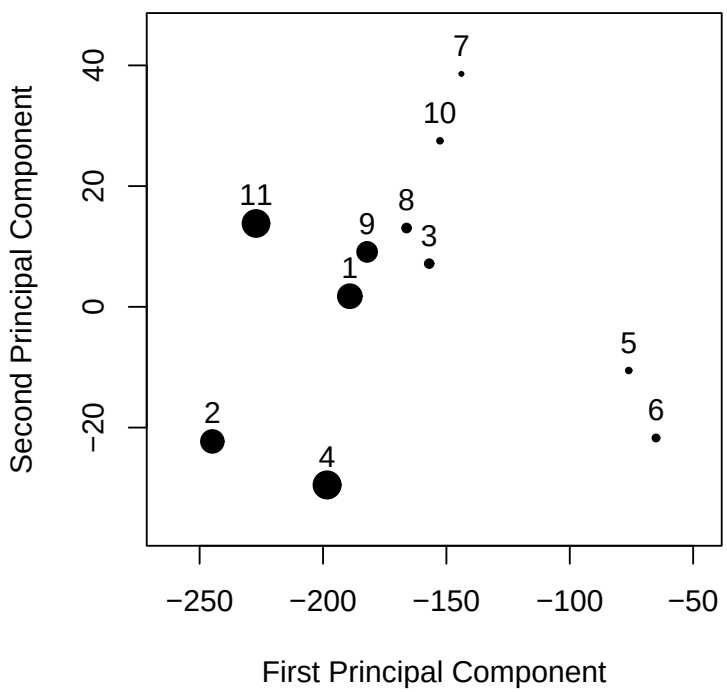
Selection for Number of Clusters

$$C(K) = \frac{\sum_{i \neq j=1}^{n} s_{ij} \tilde{s}_{ij}^{(K)}}{\sqrt{\sum_{i \neq j=1}^{n} s_{ij}^2} \sqrt{\sum_{i \neq j=1}^{n} \tilde{s}_{ij}^{(K)^2}}}$$

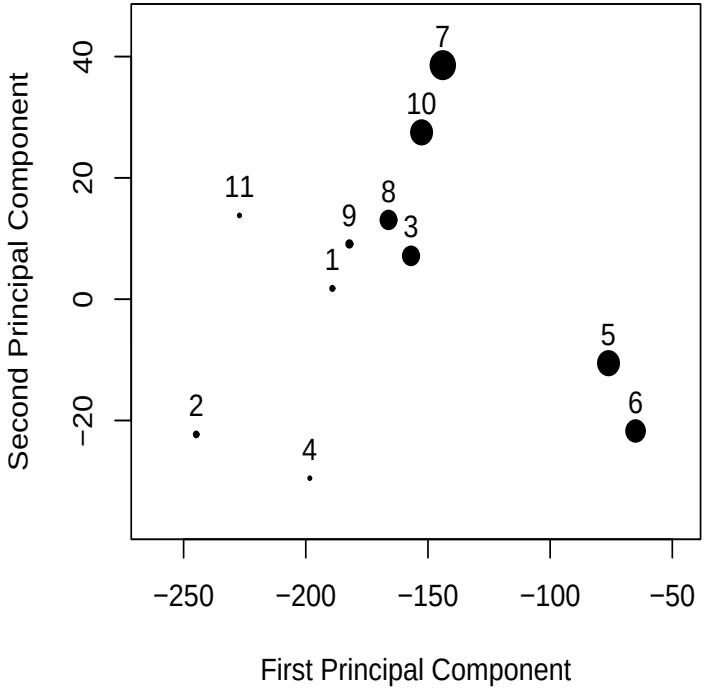
$$\tilde{s}_{ij}^{(K)} = \sum_{k=1}^{K} \sum_{l=1}^{K} w_{kl}^{(K)} u_{ik}^{(K)} u_{jl}^{(K)}, \quad i, j = 1, \ldots, n$$

K (Number of Clusters)	2	3	4
C(K)	0.93	0.90	0.91

Result of Proposed PCA

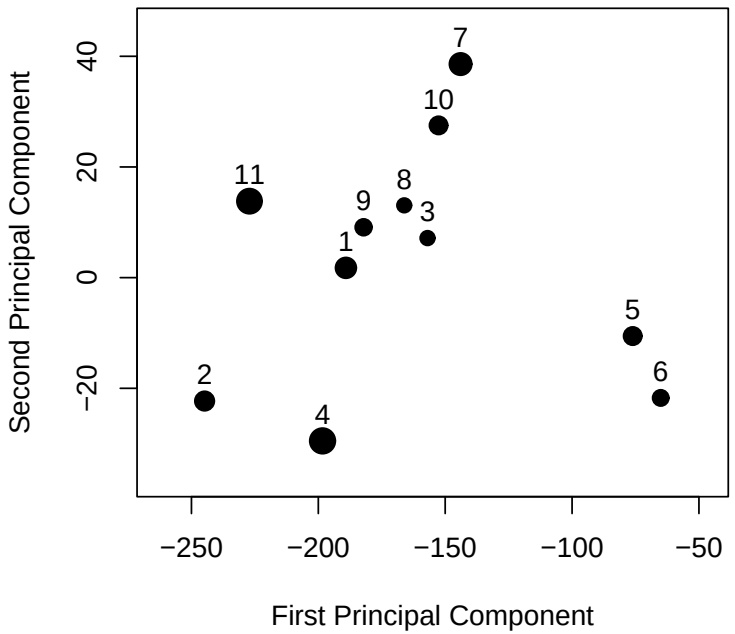


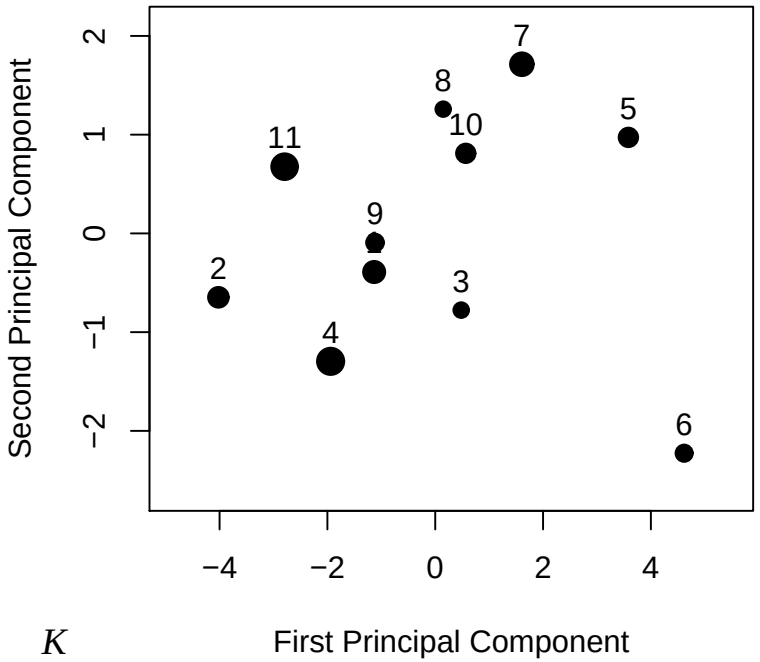
Result of Cluster 1



Result of Cluster 2

Results of Weights





Proposed PCA

Fuzzy Clustering

$$w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m}$$

$$m \in (1, \infty)$$

Centers Method

Hard Clustering

Comparison of Cumulative Proportion

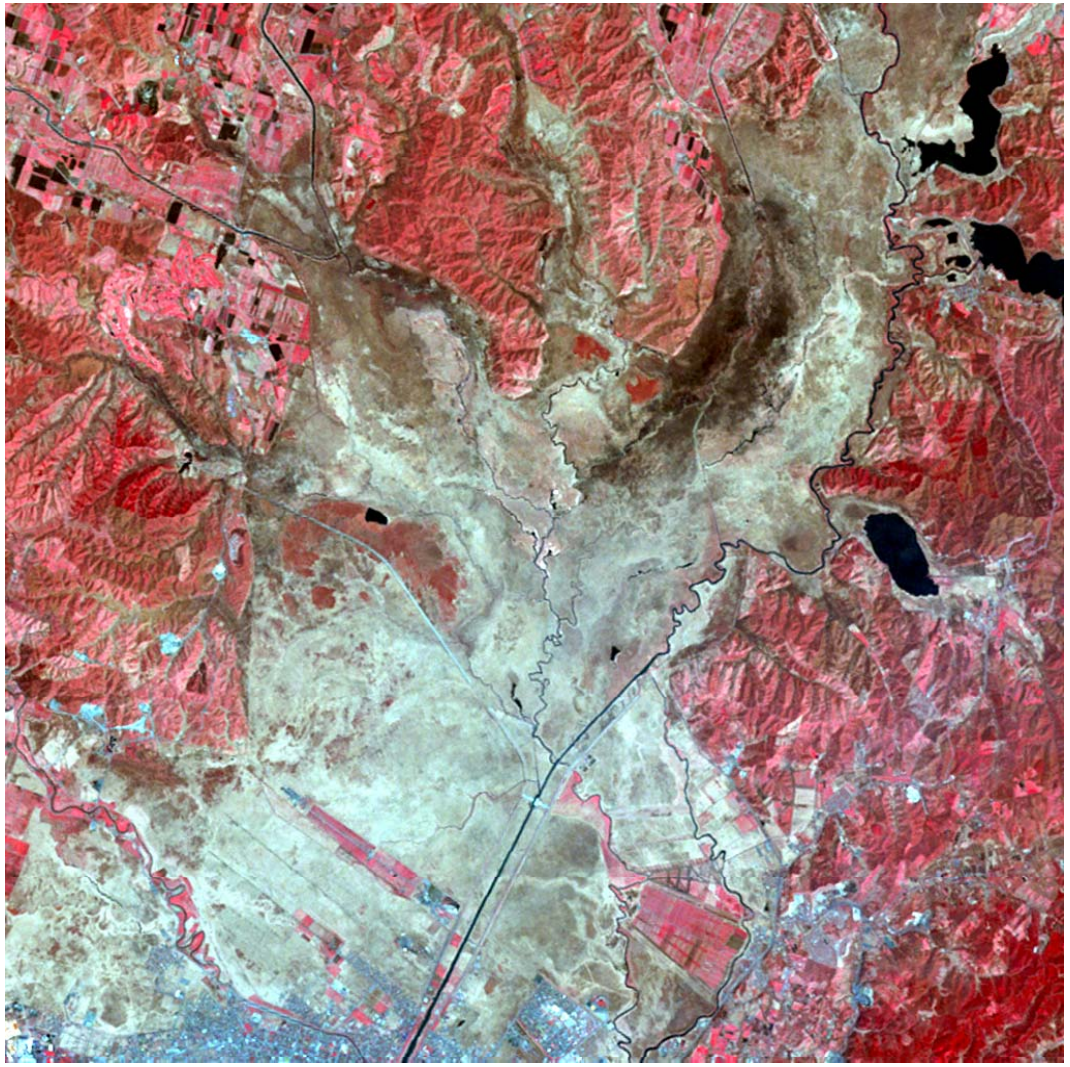
Proposed PCA	Ordinary PCA (Centers Method)
0.86	0.82

Conclusions

(1) Propose a PCA based on Fuzzy Clustering Considering Dissimilarity Structure in Higher Dimensional Space

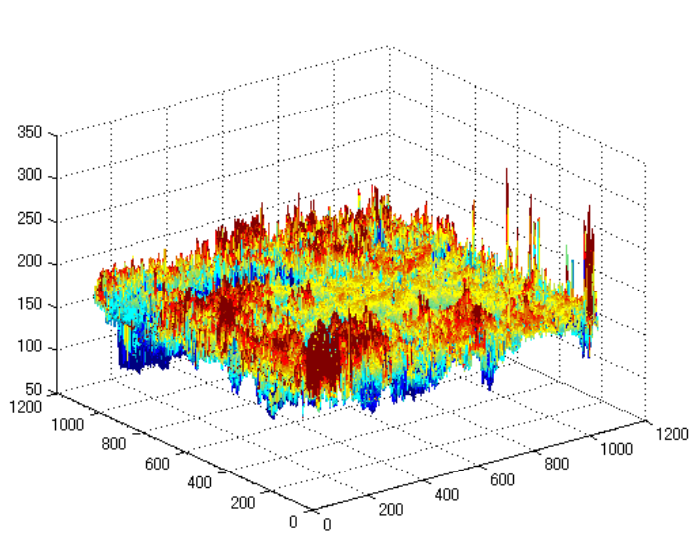
(2) Numerical Examples

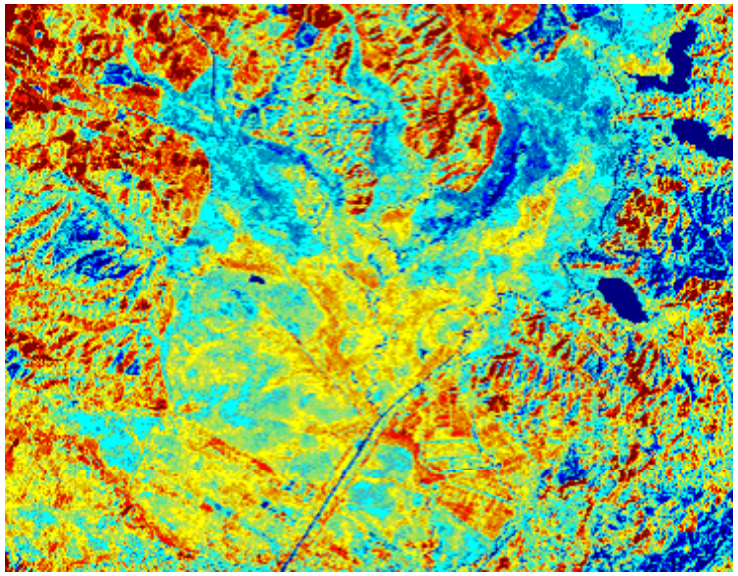
Kushiro-Marshland



Landsat Data; 1024 X 1024 pixels, 7 Kinds of Lights, July - October, 1993

Result of Proposed PCA for the First Component



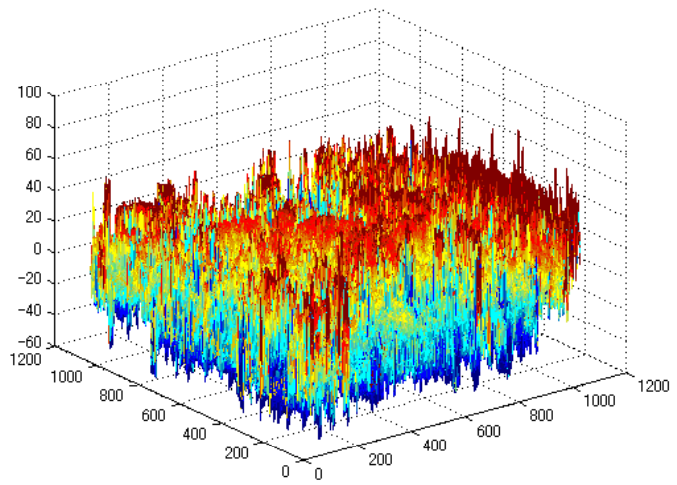


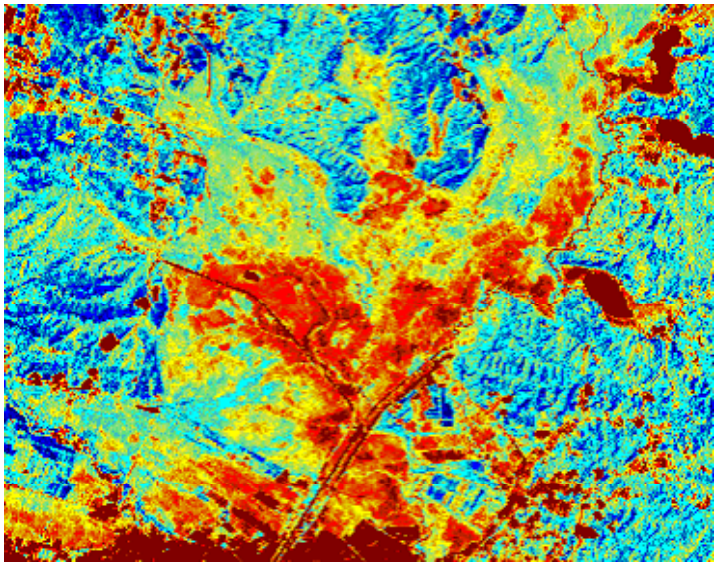
$$\hat{C} = (\hat{c}_{ab}), \quad \hat{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y_a - \bar{y}_a)(y_b - \bar{y}_b)\,\tilde{f}(y_a, y_b)\,dy_a dy_b$$

$$\tilde{f}(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i I_i(y_a, y_b)}{\|Z(i)\|} : \text{Weighted Empirical Joint Function}$$

$$w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m}, \quad m \in (1, \infty)$$

Result of Proposed PCA for the Second Component



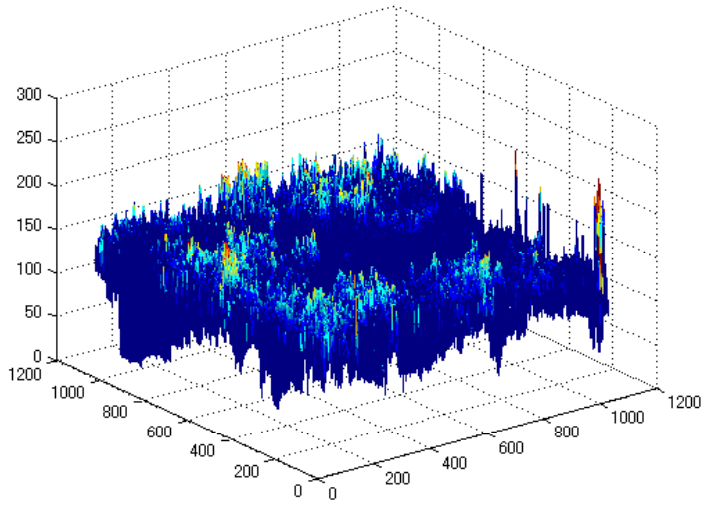


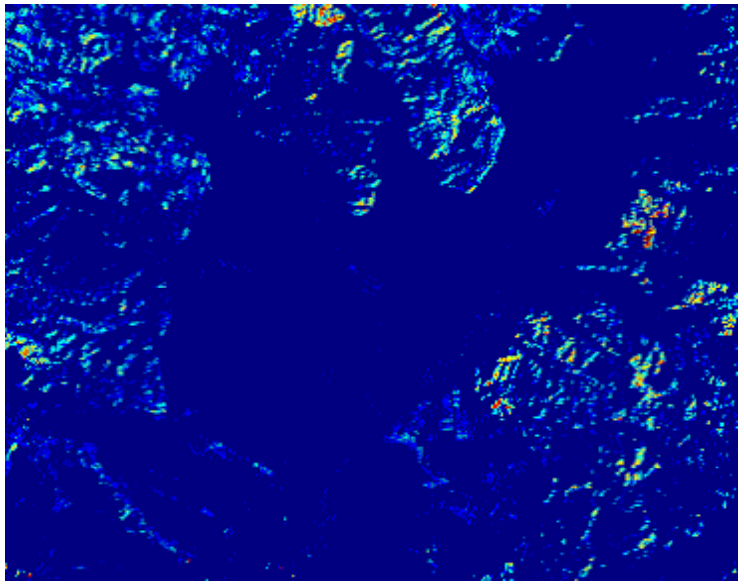
$$\hat{C} = (\hat{c}_{ab}), \quad \hat{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a - \bar{y}_a)(y_b - \bar{y}_b)\,\tilde{f}(y_a, y_b)\,dy_a dy_b$$

$$\tilde{f}(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n} \frac{w_i I_i(y_a, y_b)}{\|Z(i)\|} : \text{Weighted Empirical Joint Function}$$

$$w_i = \frac{\sum_{k=1}^{K} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m}, \quad m \in (1, \infty)$$

Result of Centers Method for the First Component



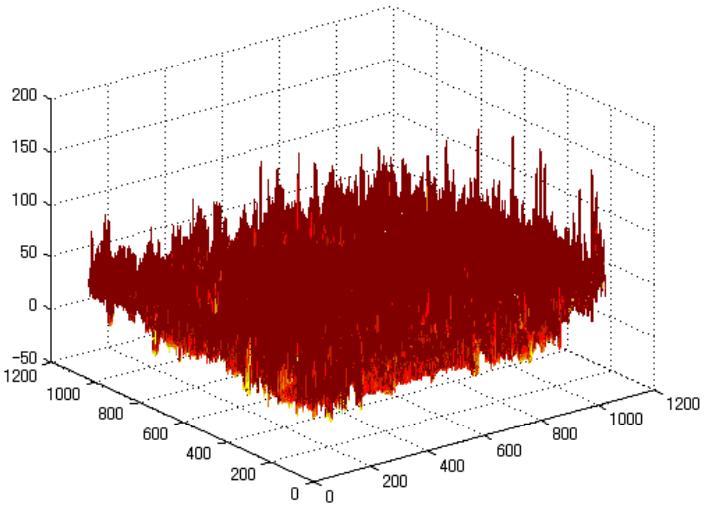


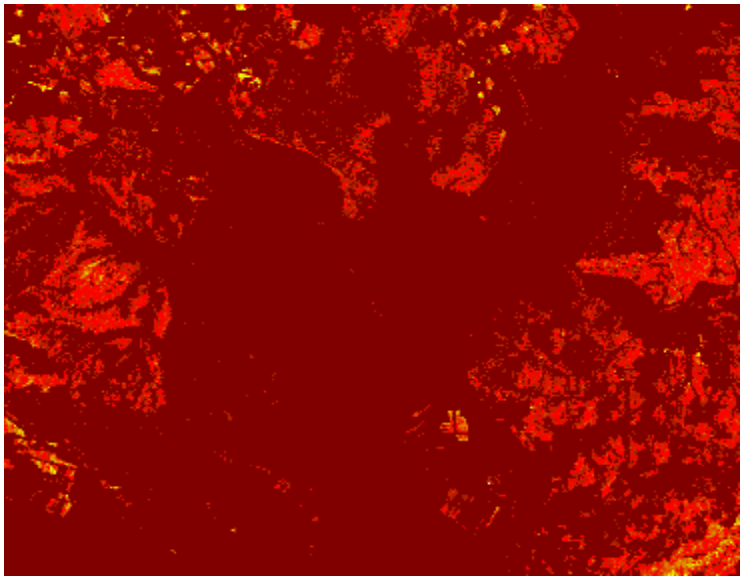
$$\tilde{C} = (\tilde{c}_{ab}), \quad \tilde{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a - \bar{y}_a)(y_b - \bar{y}_b)\, f(y_a, y_b)\, dy_a dy_b$$

$$f(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n} \frac{I_i(y_a, y_b)}{\|Z(i)\|} : \text{Empirical Joint Function}$$

$$w_i = 1, \quad \forall i$$

Result of Centers Method for the Second Component





$$\tilde{C} = (\tilde{c}_{ab}), \quad \tilde{c}_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a - \bar{y}_a)(y_b - \bar{y}_b)\, f(y_a, y_b)\, dy_a dy_b$$

$$f(y_a, y_b) = \frac{1}{n}\sum_{i=1}^{n} \frac{I_i(y_a, y_b)}{\|Z(i)\|} : \text{Empirical Joint Function}$$

$$w_i = 1, \quad \forall i$$