Symbolic Data in the Statistical Framework

L. Billard

Department of Statistics
University of Georgia
lynne@stat.uga.edu

4th International Summer School
Split, Croatia
September 9-13, 2019

The Present Too dense to 'read'

i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}	Y_{17}	Y_{18}	Y_{19}	Y_{20}	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}	Y_{26}	Y_{27}	Y_{28}	Y_{29}	Y_{30}
1	Boston	M	24	M	S	2	2	0	165	68	120	79	183	83	86	2.21	88	92	16	1.4	12	21	6.9	5.2	14.2	43.6	2.32	N	N	0
2	Boston	M	56	M	M	1	2	2	186	84	130	90	164	64	60	2.55	69	101	16	0.8	20	22	7.0	4.6	13.5	39.9	2.44	N	N	1
3	Chicago	D	48	M	M	1	3	2	175	73	126	82	229	109	122	2.10	114	80	17	1.4	13	24	7.7	4.9	14.1	44.2	2.73	Y	N	1
4	El Paso	M	47	F	M	0	1	1	141	78	121	86	239	69	74	3.45	44	90	15	1.1	14	20	6.7	4.6	13.9	40.7	2.17	Y	0	0
5	Byron	D	79	F	M	0	3	4	152	84	150	88	187	67	64	2.79	72	103	18	0.9	20	27	7.3	4.8	11.6	36.1	3.05	N	0	0
6	Concord	M	12	M	S	2	1	0	73	69	126	85	109	98	107	1.11	105	108	14	0.8	18	17	6.2	4.3	12.2	36.0	1.79	N	0	0
7	Atlanta	M	67	F	M	1	6	0	166	81	134	89	190	90	96	2.12	95	91	17	1.0	17	24	7.2	4.6	13.4	42.3	2.65	Y	6	0
8	Boston	O	73	F	M	0	2	4	164	77	121	81	181	81	84	2.24	86	112	19	0.9	22	29	8.0	4.0	14.9	43.6	3.32	N	0	0
9	Lindfield	D	29	M	M	2	0	2	227	62	124	81	214	94	101	2.28	99	89	18	1.0	18	27	7.8	4.7	15.0	43.4	3.13	N	0	0
10	Lindfield	D	44	M	M	1	3	3	216	71	125	79	218	98	107	2.23	103	83	18	1.0	18	27	7.8	4.5	12.4	37.1	3.12	Y	N	2
11	Boston	D	54	M	S	1	5	0	213	57	118	88	189	69	66	2.75	74	100	28	0.3	90	53	11.5	4.3	14.8	42.6	6.32	N	N	0
12	Chicago	M	12	F	S	2	2	0	75	69	115	81	153	54	45	2.83	58	119	20	1.0	19	31	8.3	4.4	14.3	40.7	3.59	N	0	0
13	Macon	M	73	F	M	0	3	1	152	58	123	82	188	87	93	2.15	93	69	16	1.2	13	21	6.9	4.6	12.9	37.1	2.35	N	0	0
14	Boston	D	48	M	M	0	2	4	206	73	113	72	264	72	62	3.69	49	91	14	1.2	11	16	6.1	5.0	12.9	40.5	1.67	N	N	0
15	Peoria	O	79	F	M	0	3	3	153	72	106	78	118	40	35	2.95	23	82	19	0.9	20	30	8.1	4.1	13.6	43.3	3.40	N	0	0
16	Concord	D	20	M	S	2	0	1	268	79	123	80	205	85	89	2.40	90	71	19	1.3	14	28	7.9	4.2	13.5	39.4	3.21	N	N	0
17	Boston	D	20	F	S	2	4	0	157	75	116	87	180	60	52	3.01	65	101	17	1.0	16	23	7.2	5.1	13.0	40.8	2.61	N	0	0
18	Chicago	D	17	M	S	2	2	0	161	69	114	78	169	49	39	3.45	54	96	17	1.0	16	23	7.2	4.2	13.1	40.7	2.61	N	N	0
19	Stowe	D	31	M	M	1	3	2	183	81	118	84	185	66	62	2.82	71	146	18	0.7	24	28	7.8	4.8	13.2	38.2	3.14	N	N	0
20	Tara	M	83	M	M	0	3	1	128	80	108	80	224	48	65	4.66	38	111	15	1.0	14	18	6.4	4.7	13.6	41.7	1.94	Y	N	3
21	Akron	M	20	M	S	1	3	0	182	68	114	76	150	51	40	2.94	55	58	13	1.2	11	14	5.8	4.0	13.7	40.7	1.43	N	N	0
22	Detroit	M	85	F	M	0	3	2	161	73	122	76	185	83	89	2.19	90	96	8	0.9	10	17	3.9	4.2	13.1	36.8	7.19	N	0	0
23	Ila	D	66	F	S	0	4	3	166	66	126	87	218	98	108	2.22	103	85	18	1.4	13	26	7.6	4.2	13.4	38.0	2.98	N	4	0
24	Marion	M	6	M	S	2	1	0	35	72	114	76	136	52	28	2.60	41	96	16	1.2	13	20	6.8	4.5	15.4	45.2	2.25	N	N	0
25	Albany	M	24	M	M	2	1	1	177	81	111	82	149	51	39	2.96	55	72	19	0.8	24	30	8.2	3.8	14.6	45.4	3.48	N	N	0
26	Salem	D	76	M	M	0	5	2	192	77	115	73	173	53	44	3.27	58	97	17	1.3	13	23	7.2	4.6	12.4	37.1	2.60	N	N	0
27	Quincy	O	57	M	S	1	3	2	159	72	114	75	234	131	157	1.78	139	88	17	0.8	22	24	7.2	5.0	13.0	37.5	2.65	N	N	0
28	Yuma	M	11	F	S	2	2	0	73	62	118	80	96	56	43	1.71	56	136	20	0.8	25	31	7.4	4.6	13.6	41.0	3.59	N	0	0
29	Fargo	M	27	F	M	2	2	1	124	70	114	72	167	67	63	2.49	72	104	13	0.7	20	15	5.9	4.7	13.7	41.4	1.53	N	0	0
30	Reno	D	43	F	M	2	4	4	148	66	135	97	172	52	43	3.31	57	82	17	0.8	21	25	8.3	4.2	11.7	34.0	2.83	Y	3	0
31	Amherst	M	53	F	S	1	0	3	165	65	135	96	236	134	161	1.76	141	102	9	1.0	9	4	7.0	4.6	11.2	33.9	2.18	N	1	0
32	Boston	M	14	M	S	2	1	0	132	66	125	87	149	51	39	2.96	54	120	20	1.1	18	32	8.4	5.3	12.1	38.7	3.67	N	N	0
33	New Haven	D	29	F	M	1	0	1	153	70	133	92	217	97	106	2.23	102	99	25	1.1	23	45	10.3	4.6	14.3	41.4	5.29	N	0	0
34	Kent	M	84	M	M	0	4	1	239	85	114	75	229	126	150	1.81	134	113	9	1.1	8	4	4.3	4.8	12.3	37.6	2.43	Y	N	5
35	Shelton	M	52	M	M	0	4	1	206	63	125	86	236	134	161	1.77	14	114	16	1.2	13	22	4.4	4.4	12.9	37.0	0.15	N	N	0
36	Atlanta	O	86	M	M	0	3	3	184	72	114	72	152	53	42	2.88	57	92	18	1.2	15	27	6.6	4.2	15.7	49.9	3.11	N	N	0
37	Medford	M	23	F	S	2	1	0	138	71	125	85	197	96	105	2.05	102	70	20	1.4	15	33	8.6	4.1	14.4	40.7	3.78	N	0	0
38	Bangor	M	51	M	M	2	2	2	172	81	119	78	172	73	71	2.38	77	105	17	1.3	13	23	7.2	4.8	13.5	39.3	2.60	N	N	0
39	Boston	M	70	M	M	1	6	3	183	75	114	74	151	52	42	2.90	56	79	14	0.7	21	17	6.3	5.2	12.3	36.5	1.86	Y	N	2
40	Barry	M	65	M	M	0	4	2	191	84	120	80	175	75	75	2.34	80	139	21	1.3	16	34	8.8	4.5	14.6	43.8	3.97	N	N	0
41	Trenton	M	82	M	M	0	3	4	201	79	123	84	188	87	93	2.15	93	111	18	0.8	21	26	7.5	4.6	14.5	44.6	2.91	N	N	0
42	Concord	M	60	M	S	0	4	0	175	74	117	76	163	63	58	2.58	68	112	12	1.2	9	10	5.2	4.7	13.4	38.7	0.91	N	N	0
43	Chicago	M	48	M	M	1	4	1	187	88	132	98	182	82	86	2.23	87	95	18	0.8	23	27	7.8	4.3	12.4	37.3	3.13	N	N	0
44	Omaha	M	29	M	M	1	1	1	166	59	122	82	178	78	79	2.29	83	77	22	1.5	15	38	9.3	4.3	13.5	38.9	4.43	N	N	0
45	Tampa	M	21	F	M	2	2	1	124	72	119	79	169	70	67	2.43	74	103	14	1.0	14	15	6.0	4.9	13.5	43.1	1.60	N	0	0
46	Lynn	M	81	F	M	0	5	3	161	79	128	89	210	109	124	1.93	115	86	15	0.6	23	19	7.8	4.5	11.3	32.9	2.06	Y	3	0
47	Quincy	D	70	F	M	0	3	2	178	72	119	78	230	110	124	2.09	115	87	12	1.2	10	10	5.3	4.4	12.5	39.1	8.88	Y	1	1
48	Wells	M	27	F	M	2	0	0	113	77	121	80	179	79	80	2.27	84	101	6	1.0	6	8	3.1	4.2	14.6	40.9	2.99	N	0	0
49	Buffalo	M	56	F	M	2	4	1	129	76	119	81	172	72	71	2.38	77	99	14	1.4	10	16	6.1	4.7	13.9	40.3	1.69	Y	0	4
50	Quincy	M	64	M	S	1	2	0	194	81	128	89	210	109	124	1.93	115	88	19	1.1	17	28	7.9	5.0	15.3	49.1	3.23	N	N	0
51	Boston	M	87	F	W	0	5	2	157	88	128	88	171	71	70	2.40	76	87	19	1.0	19	28	7.9	4.6	13.9	44.4	7.14	N	0	0
.																														
.																														
.																														

Types of Data:
Classical Data Value X:
- Single point in p-dimensional space
E.g., $X = 17$, $\mathbf{X} = (2.1, 18)$, $X =$ blue, $X =$ cancer (Yes, No)

Symbolic Data Value Y:
- Hypercube or **Cartesian product of distributions** in $\mathbb{R}^p$
I.e. Y = list, interval, modal in structure

Modal data:

- **Histogram,**
- empirical distribution function,
- probability distribution,
- model, ...

Weights:

- **Relative frequencies,**
- capacities,
- credibilities,
- necessities,
- possibilities, ...

Types of Data:
Classical Data Value X:
- Single point in p-dimensional space
E.g., $X = 17$, $\mathbf{X} = (2.1, 18)$, $X =$ blue, $X =$ cancer (Yes, No)

Symbolic Data Value Y:
- Hypercube or **Cartesian product of distributions** in $\mathbb{R}^p$
I.e. Y = list, interval, modal in structure

Modal data:

- **Histogram,**
- empirical distribution function,
- probability distribution,
- model, ...

Weights:

- **Relative frequencies,**
- capacities,
- credibilities,
- necessities,
- possibilities, ...

Books: Bock and Diday **(2000)**, Billard and Diday **(2006, 2019)**,
Reviews: Billard **(2011, 2014)**, Noirhomme and Brito **(2011)**, Diday **(2016)**, ..., Diday **(1987)**

Symbolic Data

How do symbolic data arise?

1. Aggregated data:
   - Research interest: classes or groups
     - age $\times$ gender categories: 35-year-old females, etc.
     - types of cancers: $\mathcal{Y} = \{\text{breast, liver, bone, } \ldots, \}$
     - auto-insurance claims: type-model of vehicle, 21-year-old males, etc.
     - patient pathway

Pathways

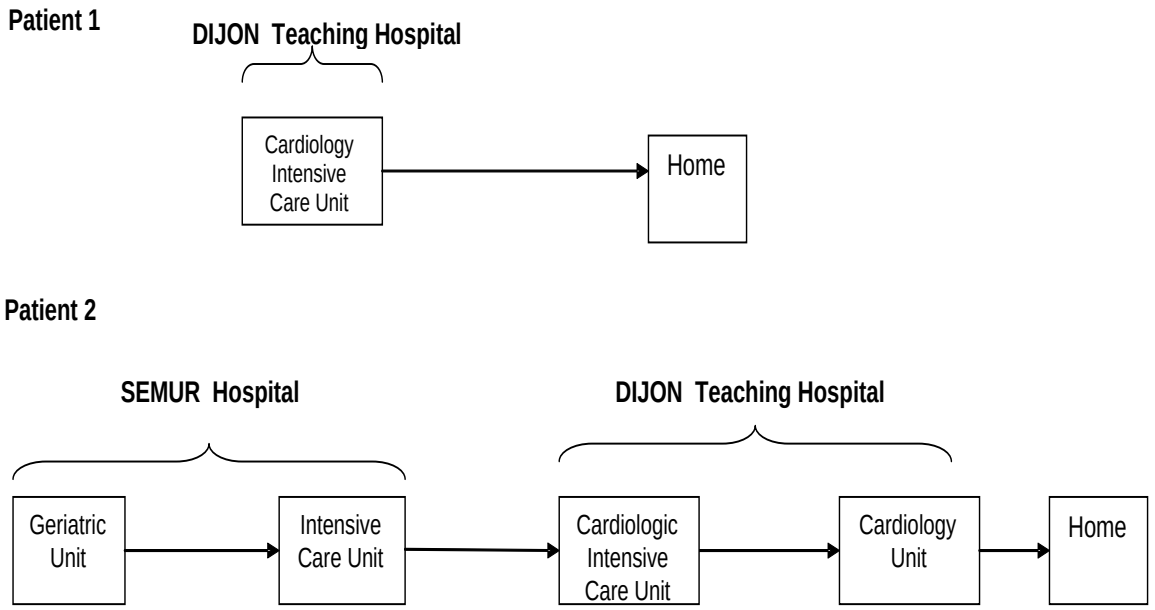
Pathways – The Scientific Question:

Does hospital pathway affect (one-year) survival rate of patients presenting with acute myocardial infarction (AMI)?

Pathways

Pathways – The Scientific Question:

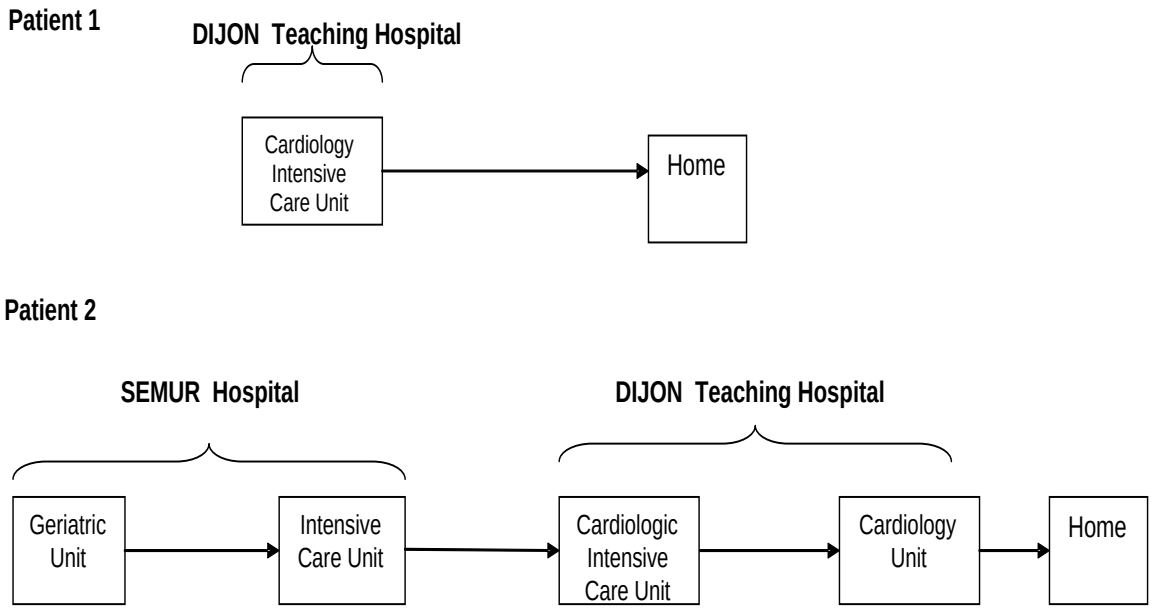
Does hospital pathway affect (one-year) survival rate of patients presenting with acute myocardial infarction (AMI)?



Possible hospital pathways for treatment of first AMI

Pathways

Pathways – The Scientific Question:
Does hospital pathway affect (one-year) survival rate of patients presenting with acute myocardial infarction (AMI)?



Possible hospital pathways for treatment of first AMI

- Patient 1: One only hospitalization
- Patient 2: Two hospitalizations in consecutive health care centers.

Patient Records - E.g. Hospitals, Cardiology

Patient	Hospital	Age	Smoker
Patient1	Hospital1	74	heavy
Patient2	Hospital1	78	light
Patient3	Hospital2	69	no
Patient4	Hospital2	73	heavy
Patient5	Hospital2	80	light
Patient6	Hospital1	70	heavy
Patient7	Hospital1	82	heavy
Patient8	Hospital3	76	no
. . .	. . .	. . .	. . .

Aggregation $\rightarrow$ Symbolic Data by Hospital $\equiv$ Pathway

Patient Records - E.g. Hospitals, Cardiology

Patient	Hospital	Age	Smoker
Patient1	Hospital1	74	heavy
Patient2	Hospital1	78	light
Patient3	Hospital2	69	no
Patient4	Hospital2	73	heavy
Patient5	Hospital2	80	light
Patient6	Hospital1	70	heavy
Patient7	Hospital1	82	heavy
Patient8	Hospital3	76	no
...	...	...	...

Aggregation $\rightarrow$ Symbolic Data by Hospital $\equiv$ Pathway

Pathway	Age	Smoker
Hospital1	[70, 82]	{light $\frac{1}{4}$, heavy $\frac{3}{4}$}
Hospital2	[69, 80]	{no, light, heavy}
Hospital3	[76, 76]	{no}
...	...	...

Patient Records - E.g. Hospitals, Cardiology

Patient	Hospital	Age	Smoker
Patient1	Hospital1	74	heavy
Patient2	Hospital1	78	light
Patient3	Hospital2	69	no
Patient4	Hospital2	73	heavy
Patient5	Hospital2	80	light
Patient6	Hospital1	70	heavy
Patient7	Hospital1	82	heavy
Patient8	Hospital3	76	no
...	...	...	...

Aggregation $\rightarrow$ Symbolic Data by Hospital $\equiv$ Pathway

Pathway	Age	Smoker
Hospital1	[70, 82]	{light $\frac{1}{4}$, heavy $\frac{3}{4}$}
Hospital2	[69, 80]	{no, light, heavy}
Hospital3	[76, 76]	{no}
...	...	...

Observations by pathways are symbolic data – obtained by aggregating classical values for patients who make up a pathway.

Patient Records - E.g. Hospitals, Cardiology

Patient	Hospital	Age	Smoker
Patient1	Hospital1	74	heavy
Patient2	Hospital1	78	light
Patient3	Hospital2	69	no
Patient4	Hospital2	73	heavy
Patient5	Hospital2	80	light
Patient6	Hospital1	70	heavy
Patient7	Hospital1	82	heavy
Patient8	Hospital3	76	no
...	...	...	...

Aggregation $\rightarrow$ Symbolic Data by Hospital $\equiv$ Pathway

Pathway	Age	Smoker
Hospital1	[70, 82]	{light $\frac{1}{4}$, heavy $\frac{3}{4}$}
Hospital2	[69, 80]	{no, light, heavy}
Hospital3	[76, 76]	{no}
...	...	...

Classical observations are special cases, as in Hospital 3 / Patient 8

Quantin et al. **(2011)**

Symbolic Data

How do symbolic data arise?

1. Aggregated data:
   - Research interest: classes or groups
     - age $\times$ gender categories: 35-year-old females, etc.
     - types of cancers: $\mathcal{Y}$ = {breast, liver, bone, . . . , }
     - auto-insurance claims: type-model of vehicle, 21-year-old males, etc.
     - patient pathway
     - . . . the list is endless . . .
2. Naturally occurring symbolic data:
   - Species: E.g., Pileus cap width (*arorae*) = [3.0, 8.0].
   - Pulse rate: 64 $\pm$ 2=[62, 66].
   - Daily temperature: [47, 70].
3. Government data: census data.
4. Confidentialities: Salary = [30K - δ_1, 30K + δ_2], $\delta_1 \neq \delta_2$.
5. .... and so on .....

Multi-valued variables, Lists

Y = Bird Colors

u	Bird	Major Colors
w_1	Magpie	{black, white}
w_2	Kookaburra	{brown, black, white, blue}
w_3	Galah	{pink, grey}
w_4	Cardinal	{red, black}
w_5	Goldfinch	{black, yellow}
w_6	Quetzal	{red, green, white}
w_7	Toucan	{ black, yellow, red, green}
w_8	Rainbow Lorikeet	{blue, yellow, green, red, violet, orange}

Multi-valued/Lists Data

Multi-valued variables, Lists

Rainbow Lorikeet

Interval-valued variables

- Species: E.g., Pileus cap width (*arorae*) = [3.0, 8.0].
- Pulse rate: 64 ± 2=[62, 66].
- Daily temperature: [65, 82].

Histogram-valued variables

E.g., Diagnostics of flights into JFK

	Y_1=Flight Time			Y_2 = Arrival Delay			Y_3 = Departure Delay			Y_4 = Weather Delay	
Airline	< 120	$[120, 220]$	> 220	< 0	$[0, 60]$	> 60	< 0	$[0, 60]$	> 60	No	Yes
1	.15	.62	.23	.42	.46	.12	.44	.47	.09	.92	.08
2	.89	.11	.00	.52	.39	.09	.32	.60	.08	.90	.10
...	...	...	...	...	...	...	...	...	...	...	...

Mixed-valued variables

Species	**DNA Sequence**	Y_1	Y_2	Y_3	Y_4
D. pseudsobscura	ACCGTCCGTTA	{7, 8}	Long	[5.4, 6.7]	[0.29, 0.33]
D. obscura	ACAGGCCGTGA	{5, 7}	Medium	[9.0, 11.2]	[0.43, 0.49]
D. melanogaster	AACGTCCGTGC	{3, 4}	Short	[16.3, 21.1]	[0.50, 0.89]

Distinctive Features - 1. Data Structures

1. Symbolic data analyses take into account data structures

E.g., Symbolic data have internal variation

Consider	Y = weight	$n = 1$
Classical:	$y = 135$	$S^2 = 0$
Symbolic:	$y_1 = [132, 138]$	$S_1^2 = 3$
	$y_2 = [129, 141]$	$S_2^2 = 12$

Note $\bar{Y} = \bar{Y}_1 = \bar{Y}_2 = 135$

Distinctive Features - 1. Data Structures

1. Symbolic data analyses take into account data structures

E.g., Symbolic data have internal variation

Consider	Y = weight	$n = 1$
Classical:	$y = 135$	$S^2 = 0$
Symbolic:	$y_1 = [132, 138]$	$S_1^2 = 3$
	$y_2 = [129, 141]$	$S_2^2 = 12$

Note $\bar{Y} = \bar{Y}_1 = \bar{Y}_2 = 135$

Classical analyses on interval midpoints do not take into account the internal variations.

Can show that
Total Variation (SS/SP) = Within Variation + Between Variation

Distinctive Features - 3. Output v-v Input

3. What is the output format? – It depends!

Distinctive Features - 3. Output v-v Input

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**

Distinctive Features - 3. Output v-v Input

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical covariance is a point – Billard **(2008)**

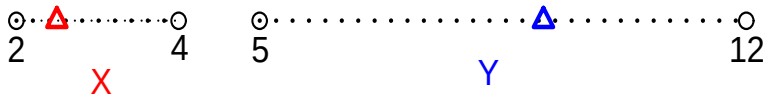
Distinctive Features - 3. Output v-v Input

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical covariance is a point – Billard **(2008)**
   - Principal components are histograms – Le-Rademacher and Billard **(2013)**

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical covariance is a point – Billard **(2008)**
   - Principal components are histograms – Le-Rademacher and Billard **(2013)**
   - Regression:

The Data are Symbolic; the Variables are Standard
Symbolic Data are NOT fuzzy data

E.g., Suppose X takes a value in $[2, 4]$ and Y takes a value in $[5, 12]$



E.g., X=2.5 with weight .8, and Y=9 with weight .6.

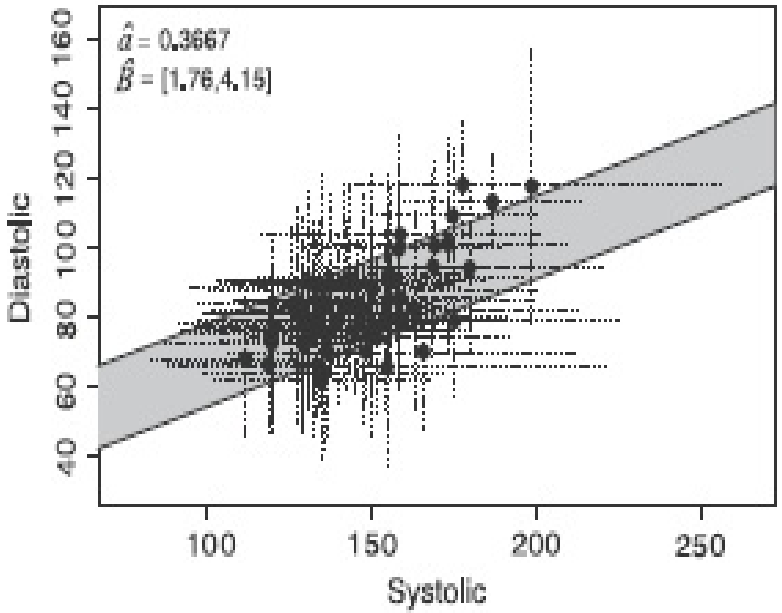
Simple regression: $Y = \beta X + \varepsilon$

Then, what is β? Suppose $n = 1$ and no error

If (X, Y) has weight (1,1), then β ranges from 1.25 to 6. I.e.,
parameter β is also fuzzy.
If (X,Y) has weight (.8,.6), then range of β = ??

Distinctive Features - Regression Output

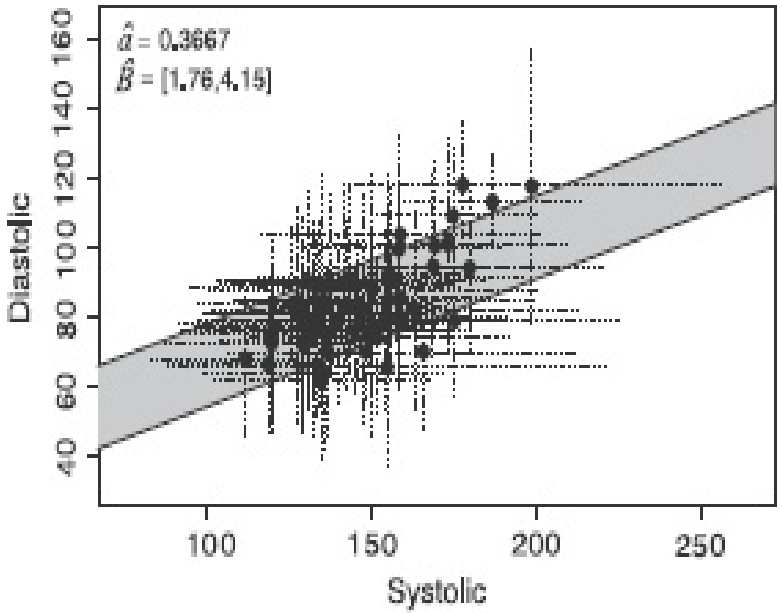
Fuzzy Regression



Sinova et al. **(2012)**

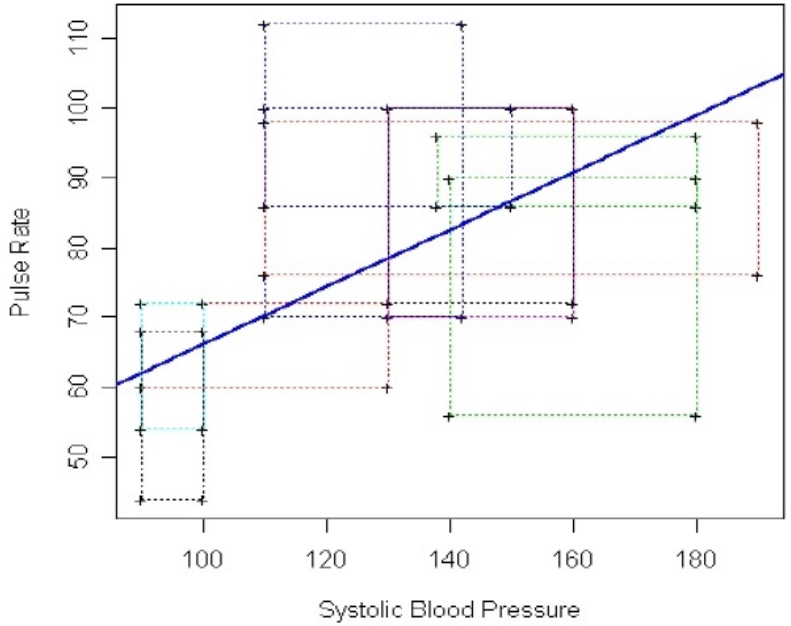
Distinctive Features - Regression Output

Fuzzy Regression



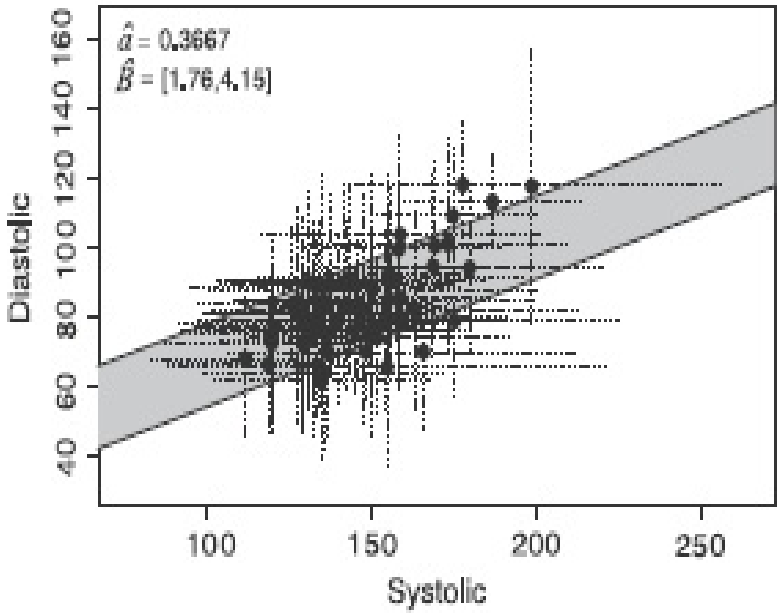
Sinova et al. **(2012)**

Symbolic Regression

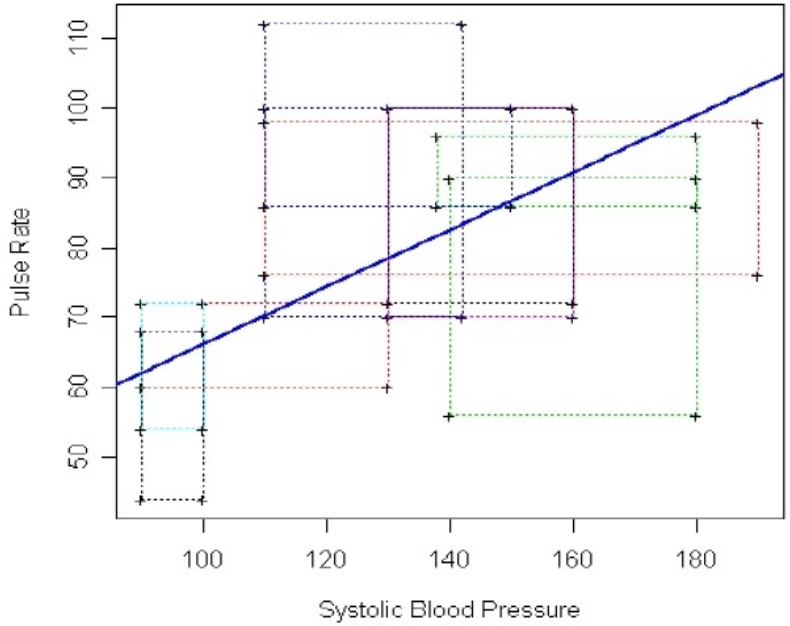


Distinctive Features - Regression Output

Fuzzy Regression



Symbolic Regression



Sinova et al. **(2012)**

Zadeh **(1965):** fuzzy logic is different in character from probability, and is not a replacement for it. Different domains

---

Use standard algebra; not interval arithmetic

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical covariance is a point – Billard **(2008)**
   - Principal components are histograms – Le-Rademacher and Billard **(2013)**
   - Regression: For model $Y = \beta_0 + \beta_1 X$
     What does it mean for slope β_1 to be an interval?
     What does it mean for intercept β_0 to be an interval?

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical covariance is a point – Billard **(2008)**
   - Principal components are histograms – Le-Rademacher and Billard **(2013)**
   - Regression: For model $Y = \beta_0 + \beta_1 X$
     What does it mean for slope β_1 to be an interval?
     What does it mean for intercept β_0 to be an interval?
   - Clustering - partitions, hierarchy, pyramids - varies
   - ...

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical covariance is a point – Billard **(2008)**
   - Principal components are histograms – Le-Rademacher and Billard **(2013)**
   - Regression: For model $Y = \beta_0 + \beta_1 X$
     What does it mean for slope β_1 to be an interval?
     What does it mean for intercept β_0 to be an interval?
   - Clustering - partitions, hierarchy, pyramids - varies
   - ...

Conclusion: it depends on methodology
Remember symbolic data are not fuzzy data; and standard arithmetic is not interval arithmetic. These are different domains

3. What is the output format? – It depends!
   Consider interval data input:
   - Sample/empirical mean is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical variance is a point – Bertrand and Goupil **(2000)**
   - Sample/empirical covariance is a point – Billard **(2008)**
   - Principal components are histograms – Le-Rademacher and Billard **(2013)**
   - Regression: For model $Y = \beta_0 + \beta_1 X$
     What does it mean for slope β_1 to be an interval?
     What does it mean for intercept β_0 to be an interval?
   - Clustering - partitions, hierarchy, pyramids - varies
   - ...

Conclusion: it depends on methodology
Remember symbolic data are not fuzzy data; and standard arithmetic is not interval arithmetic. These are different domains

Same differences occur for other types of symbolic data

Classical Analysis - PCA

Classical Principal Components (PC)

Observations: $\mathbf{X}_i = (X_{i1}, ..., X_{ip}), \quad i = 1, ..., n,$
Then, for $\nu = 1, \ldots, p$,

$$PC_\nu = e_{\nu_1} X_1 + \cdots + e_{\nu_p} X_p$$

where $\boldsymbol{\lambda}_\nu = (\lambda_{\nu_1}, \ldots, \lambda_{\nu_p}), \quad \mathbf{e}_\nu = (\mathbf{e}_{\nu_1}, \ldots, \mathbf{e}_{\nu_p})$
are ν^{th} eigenvalue and ν^{th} eigenvector, respectively, of variance-covariance matrix $^\circ$, and with $\sum_j \lambda_{\nu_j} = 1$;

$$Var(PC_\nu) = \lambda_\nu, \text{ and } Cov(PC_\nu, PC_{\nu'}) = 0, \quad \nu \neq \nu'.$$

Classical Analysis - PCA

Classical Principal Components (PC)

Observations: $\mathbf{X}_i = (X_{i1}, ..., X_{ip}), \quad i = 1, ..., n,$
Then, for $\nu = 1, \ldots, p$,

$$PC_\nu = e_{\nu_1} X_1 + \cdots + e_{\nu_p} X_p$$

where $\boldsymbol{\lambda}_\nu = (\lambda_{\nu_1}, \ldots, \lambda_{\nu_p}), \quad \mathbf{e}_\nu = (\mathbf{e}_{\nu_1}, \ldots, \mathbf{e}_{\nu_p})$
are ν^{th} eigenvalue and ν^{th} eigenvector, respectively, of variance-covariance matrix $^\circ$, and with $\sum_j \lambda_{\nu_j} = 1$;

$$Var(PC_\nu) = \lambda_\nu, \text{ and } Cov(PC_\nu, PC_{\nu'}) = 0, \quad \nu \neq \nu'.$$

Two key steps:

1. Calculate variance-covariance matrix $^\circ$
2. Project observation $\mathbf{X}_i$ onto PC space.

Interval Data – previous methods:

- Centers - Chouakria **(1998)**; Cazes et al. **(1997)**
  Ignores internal variations.
  Recall Total (SS/SP) = **B**etween (SS/SP) + **W**ithin (SS/SP)

Interval Data – previous methods:

- Centers - Chouakria **(1998)**; Cazes et al. **(1997)**
  Ignores internal variations.
  Recall Total (SS/SP) = **B**etween (SS/SP) + **W**ithin (SS/SP)
- Vertices - Chouakria **(1998)**; Cazes et al. (1997)
  "**V**ertices"(SS/SP) = **B**(SS/SP) + (SS/SP$_v$ $\neq$ **W**(SS/SP))

Interval Data – previous methods:

- Centers - Chouakria **(1998)**; Cazes et al. **(1997)**
  Ignores internal variations.
  Recall Total (SS/SP) = **B**etween (SS/SP) + **W**ithin (SS/SP)
- Vertices - Chouakria **(1998)**; Cazes et al. (1997)
  "**V**ertices"(SS/SP) = **B**(SS/SP) + (SS/SP$_v$ $\neq$ **W**(SS/SP))
- Symbolic Object (SO) - Lauro and Palumbo **(2000)**
  Uses vertices to do centers method

Interval Data – previous methods:

- Centers - Chouakria **(1998)**; Cazes et al. **(1997)**
  Ignores internal variations.
  Recall Total (SS/SP) = **B**etween (SS/SP) + **W**ithin (SS/SP)
- Vertices - Chouakria **(1998)**; Cazes et al. (1997)
  "**V**ertices"(SS/SP) = **B**(SS/SP) + (SS/SP$_v$ $\neq$ **W**(SS/SP))
- Symbolic Object (SO) - Lauro and Palumbo **(2000)**
  Uses vertices to do centers method
- Midpoints and range - Palumbo and Lauro **(2003)**
  Converts interval $[a, b]$ into midpoints $m = (a + b)/2$ and range $r = (b - a)/2$; **R**ange(SS/SP) $\neq$ **W**ithin(SS/SP)

Interval Data – previous methods:

- Centers - Chouakria **(1998)**; Cazes et al. **(1997)**
  Ignores internal variations.
  Recall Total (SS/SP) = **B**etween (SS/SP) + **W**ithin (SS/SP)
- Vertices - Chouakria **(1998)**; Cazes et al. (1997)
  "**V**ertices"(SS/SP) = **B**(SS/SP) + (SS/SP$_v$ $\neq$ **W**(SS/SP))
- Symbolic Object (SO) - Lauro and Palumbo **(2000)**
  Uses vertices to do centers method
- Midpoints and range - Palumbo and Lauro **(2003)**
  Converts interval $[a, b]$ into midpoints $m = (a + b)/2$ and range $r = (b - a)/2$; **R**ange(SS/SP) $\neq$ **W**ithin(SS/SP)
- Interval algebra - Gioia and Lauro **(2006)**
- SO, RT, Mixed (and IA) - Lauro Verde Irpino **(2008)**
- Vertices+ and Classical surrogates - Douzal-Chouakria et al. **(2011)**

All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.

All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.
Other considerations include:

1. Implicit independence assumptions:
   - Between endpoints (in vertices method)
   - Between midpoints and ranges (range methods), etc.

All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.
Other considerations include:

1. Implicit independence assumptions: Not sustainable
   - Between endpoints (in vertices method)
   - Between midpoints and ranges (range methods), etc.

All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.
Other considerations include:

1. Implicit independence assumptions: Not sustainable
   - Between endpoints (in vertices method)
   - Between midpoints and ranges (range methods), etc.
2. Intervals with common midpoints

   [9,11], [1,19], [2,18], . . .

All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.
Other considerations include:

1. Implicit independence assumptions: Not sustainable
   - Between endpoints (in vertices method)
   - Between midpoints and ranges (range methods), etc.
2. Intervals with common midpoints

   [9,11], [1,19], [2,18], . . .

   - Eigenvalues are zero for midpoint covariance matrix
   - $\Rightarrow$ Centers, SO, RT will not work

All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.
Other considerations include:

1. Implicit independence assumptions: Not sustainable
   - Between endpoints (in vertices method)
   - Between midpoints and ranges (range methods), etc.
2. Intervals with common midpoints

   [9,11], [1,19], [2,18], . . .

   - Eigenvalues are zero for midpoint covariance matrix
   - $\Rightarrow$ Centers, SO, RT will not work
3. Intervals with common ranges

   [0,10], [20,30], [120,130], . . .

All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.
Other considerations include:

1. Implicit independence assumptions: Not sustainable
   - Between endpoints (in vertices method)
   - Between midpoints and ranges (range methods), etc.
2. Intervals with common midpoints

   [9,11], [1,19], [2,18], . . .

   - Eigenvalues are zero for midpoint covariance matrix
   - $\Rightarrow$ Centers, SO, RT will not work
3. Intervals with common ranges

   [0,10], [20,30], [120,130], . . .

   - Eigenvalues are zero for range covariance matrix
   - $\Rightarrow$ RT will not work

PCA Methodology

(Back to:) All these methods fail in some way to use all the variations inherent in the data; there is a loss of information.

;

PCA Methodology

(Back to:) All these methods fail in some way to use all the variations inherent in the data; there is a loss of information. However, Billard **(2008)** shows that for
$Y_u = [a_u, b_u], \quad X_u = [c_u, d_u],$

$$\text{Total}SP = \frac{1}{6}\sum_{u \in E}[2(a_u - \bar{Y})(c_u - \bar{X}) + (a_u - \bar{Y})(d_u - \bar{X}) + (b_u - \bar{Y})(c_u - \bar{X}) + 2(b_u - \bar{Y})(d_u - \bar{X})]$$

;

PCA Methodology

(Back to:) All these methods fail in some way to use all the variations inherent in the data; there is a loss of information. However, Billard **(2008)** shows that for
$Y_u = [a_u, b_u], \quad X_u = [c_u, d_u],$

$$\text{Total}SP = \frac{1}{6}\sum_{u \in E}[2(a_u - \bar{Y})(c_u - \bar{X}) + (a_u - \bar{Y})(d_u - \bar{X}) + (b_u - \bar{Y})(c_u - \bar{X}) + 2(b_u - \bar{Y})(d_u - \bar{X})]$$

Hence,

- Symbolic covariance method - Le-Rademacher and Billard **(2012)**
  - Calculates total variations exactly $\rightarrow$ exact covariance $^\circ$
    hence $\rightarrow$ exact $\boldsymbol{\lambda}_\nu, \mathbf{e}_\nu$ and PC_ν;

PCA Methodology

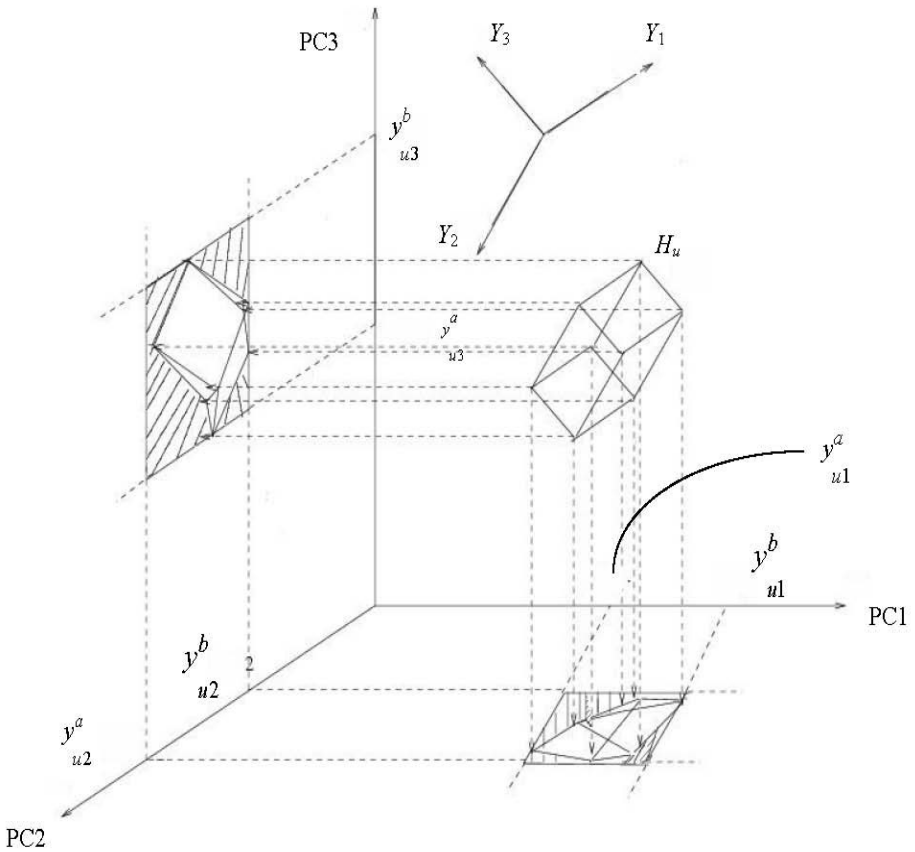
(Back to:) All these methods fail in some way to use all the variations inherent in the data; there is a loss of information. However, Billard **(2008)** shows that for
$Y_u = [a_u, b_u]$, $X_u = [c_u, d_u]$,

$$\text{Total}SP = \frac{1}{6}\sum_{u \in E}[2(a_u - \bar{Y})(c_u - \bar{X}) + (a_u - \bar{Y})(d_u - \bar{X}) + (b_u - \bar{Y})(c_u - \bar{X}) + 2(b_u - \bar{Y})(d_u - \bar{X})]$$

Hence,

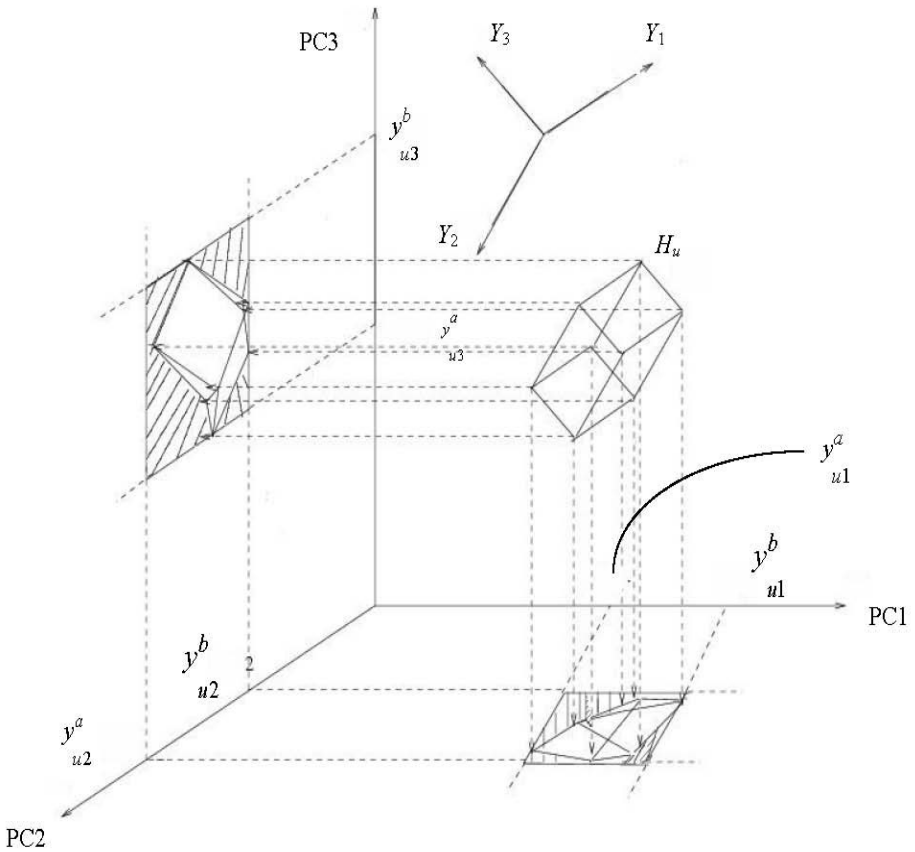
- Symbolic covariance method - Le-Rademacher and Billard **(2012)**
  - Calculates total variations exactly $\rightarrow$ exact covariance $^\circ$
    hence $\rightarrow$ exact $\boldsymbol{\lambda}_\nu, \mathbf{e}_\nu$ and PC_ν;
  - Also, a new PC space visualization - using polytope theory

Visualization?

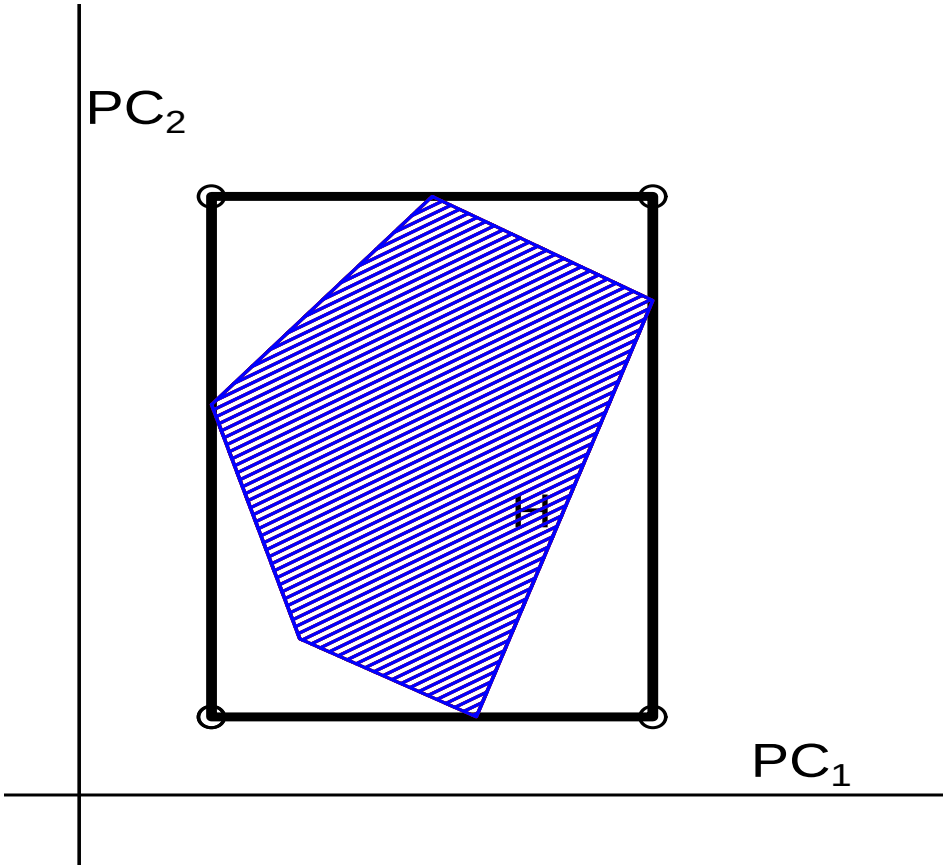


Hypercube Projection

Visualization?



Hypercube Projection



Maximum Covering Area Rectangle (MCARs)

Illustration - Faces Dataset

Faces dataset - LeRoy et al. **(1996)** - $p = 6$ variables

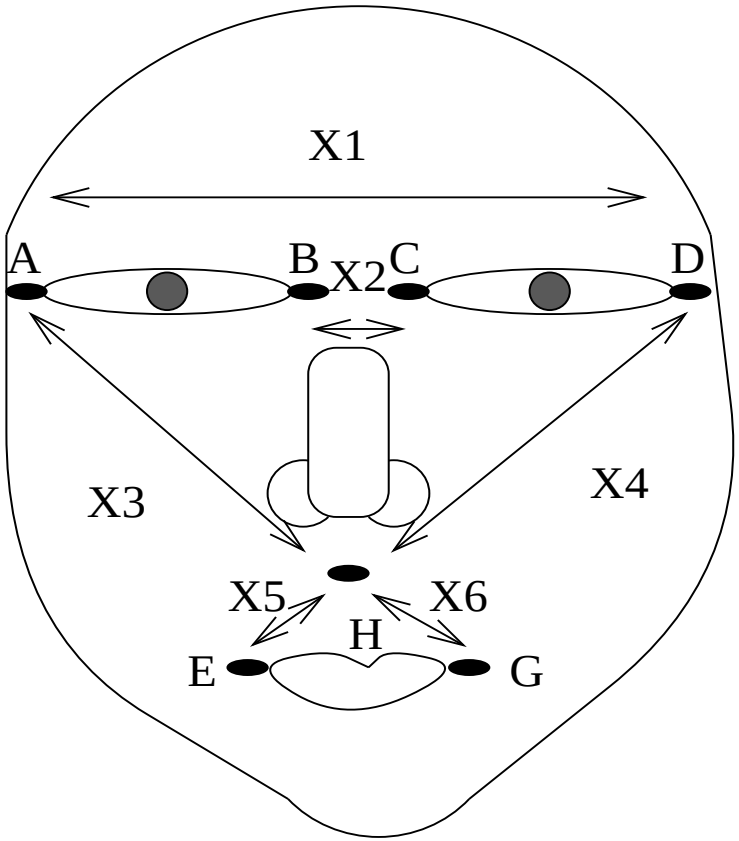
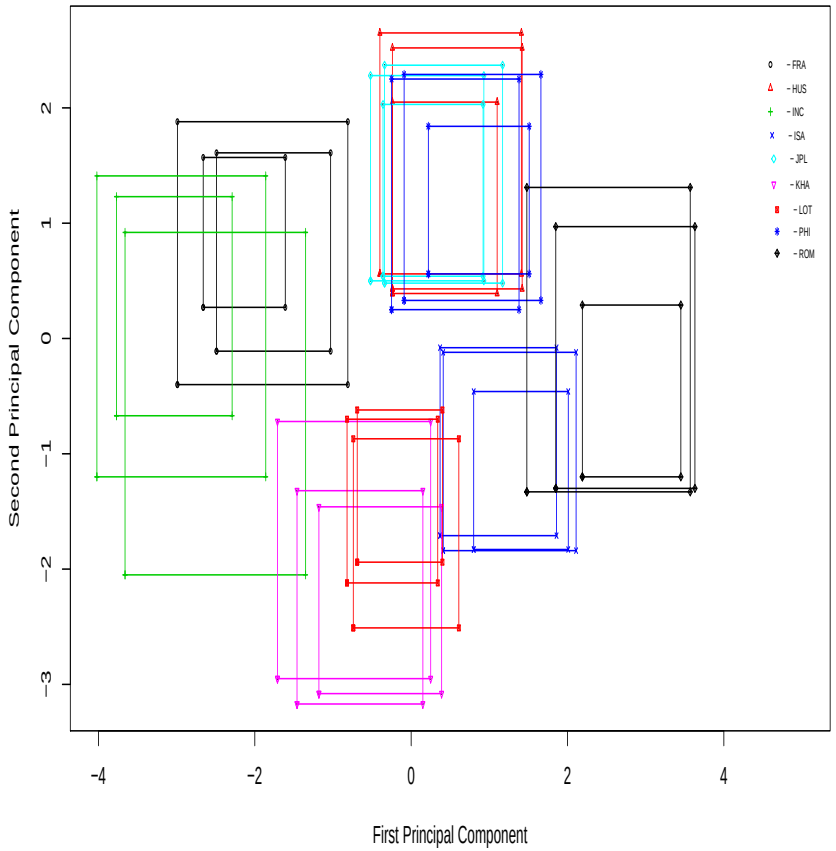


Illustration -

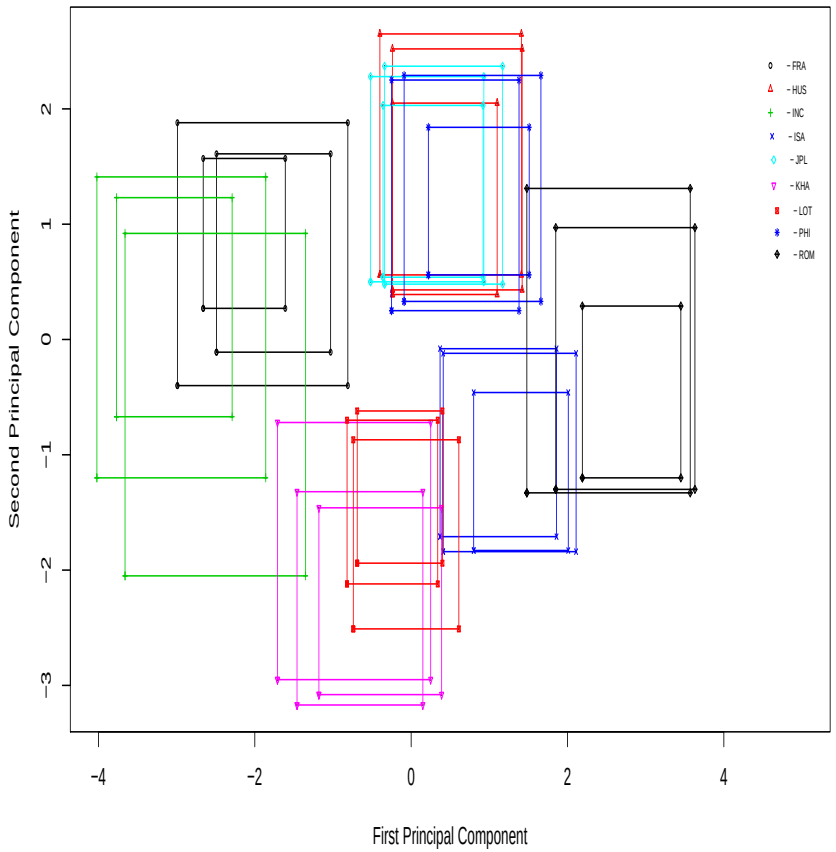
Faces dataset - 27 faces in sets of 3



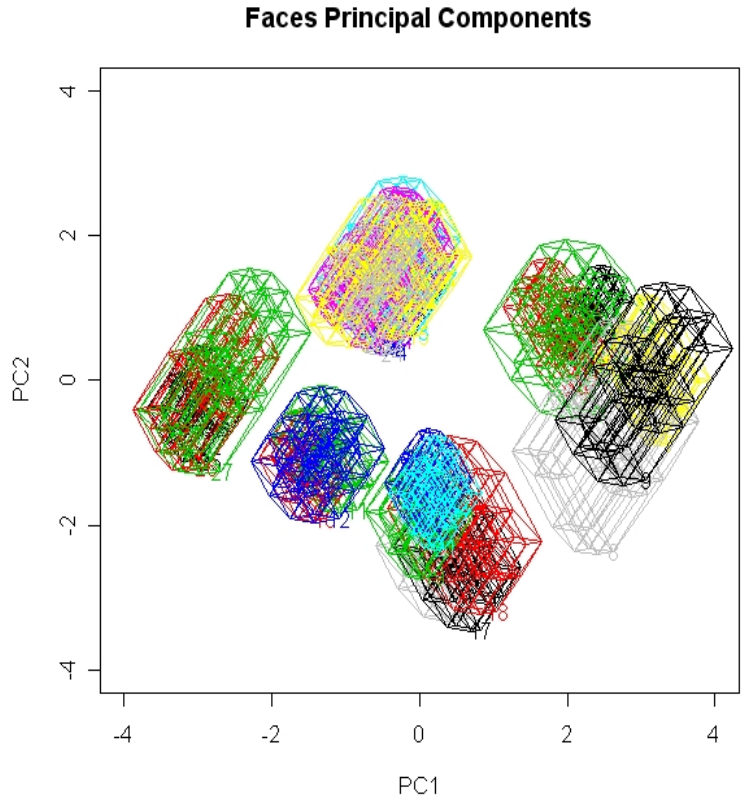
Vertices

Illustration -

Faces dataset - 27 faces in sets of 3



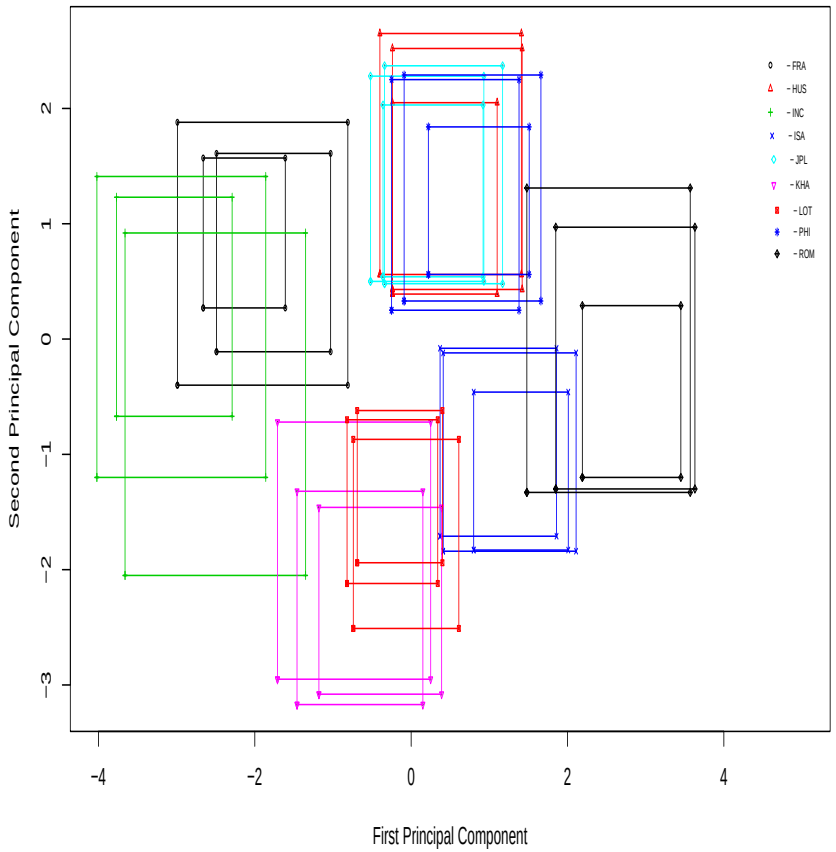
Vertices



Symbolic covariance and Polytopes

Illustration - Vertices method

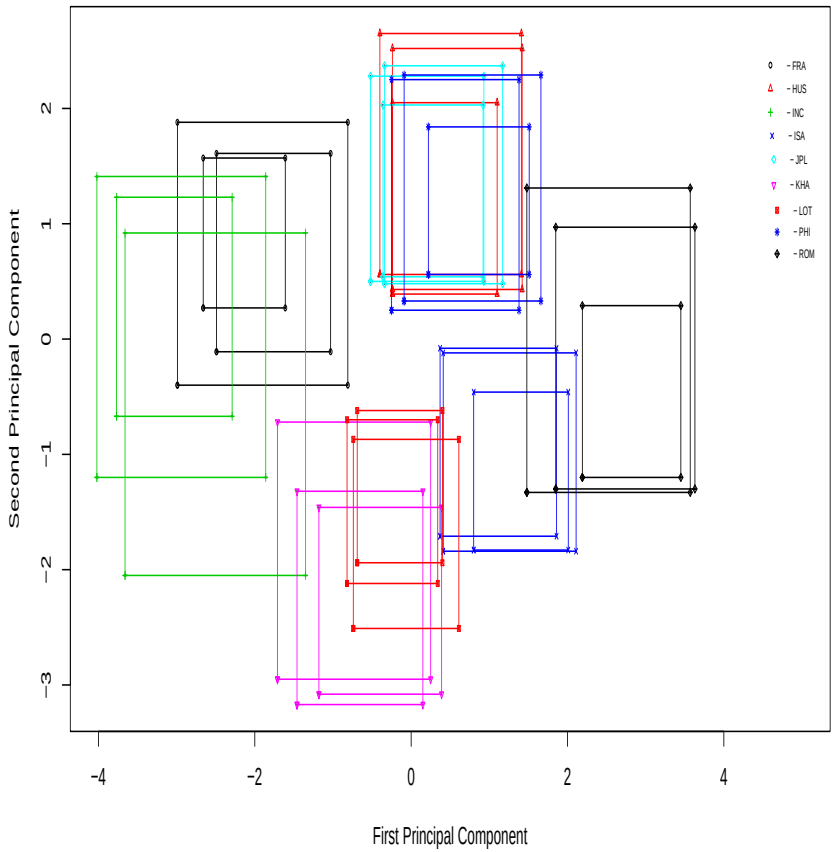
Faces dataset - 27 faces in sets of 3



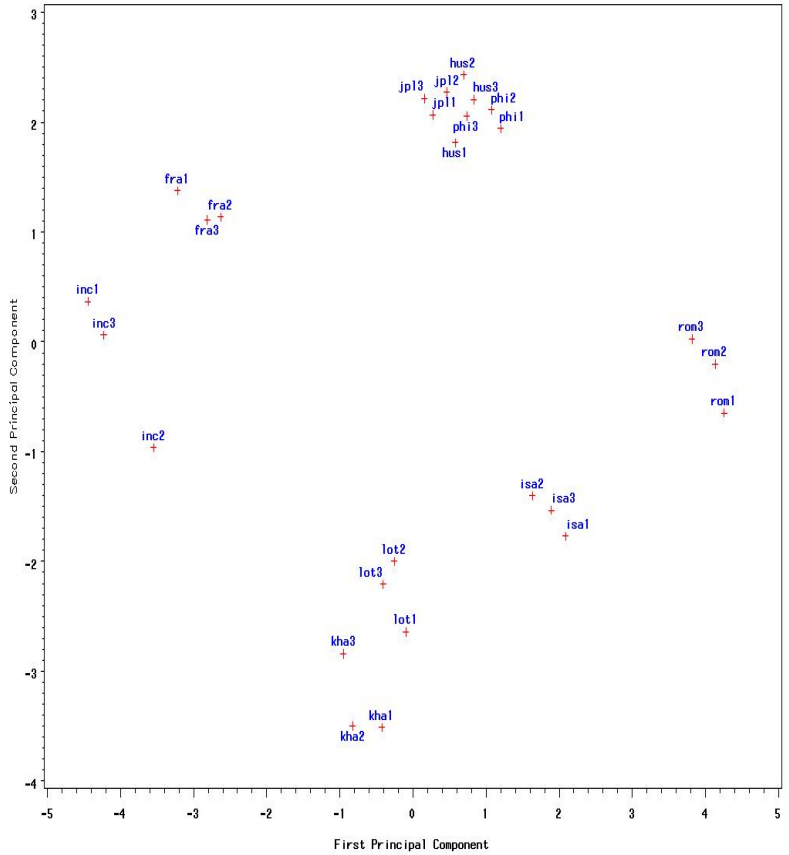
Vertices
(Cf. Centers)

Illustration - Vertices method

Faces dataset - 27 faces in sets of 3

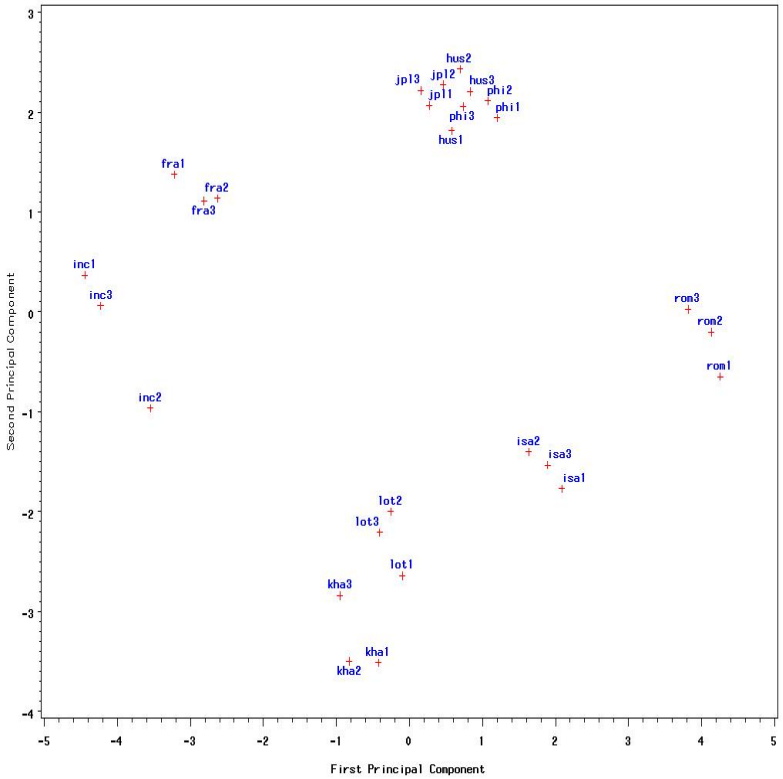


Vertices
(Cf. Centers)



Classical on Endpoints
(Cf. Classical on Midpoints)

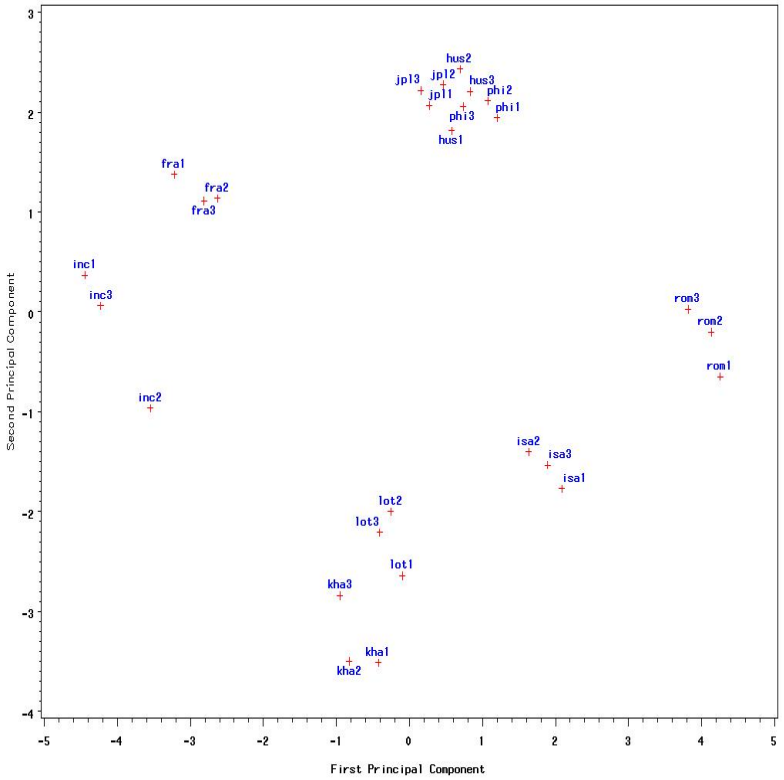
Faces dataset - 27 faces in sets of 3 – Classical Centers/Range



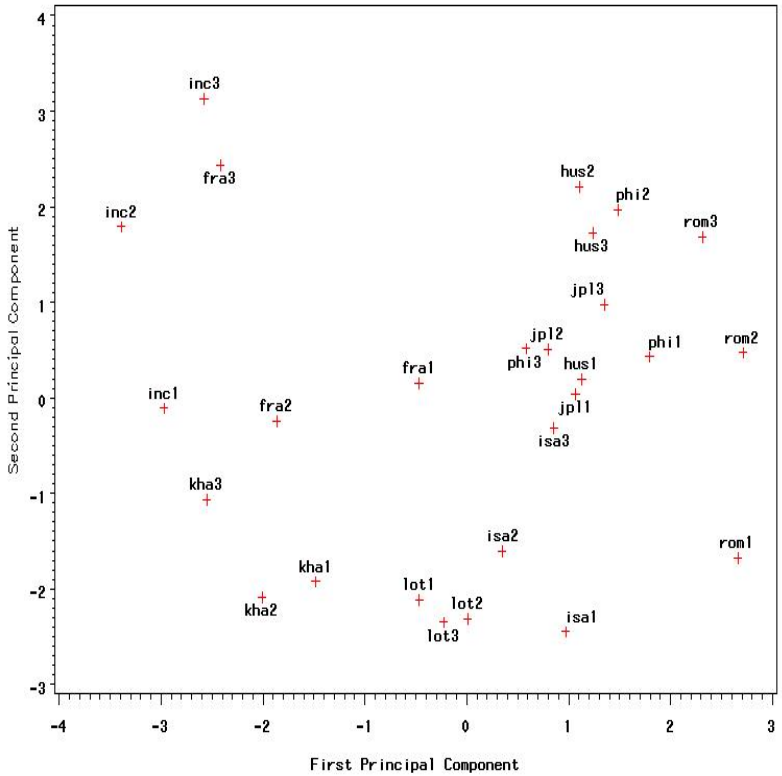
Classical on Endpoints

(Cf. Classical on Midpoints)

Faces dataset - 27 faces in sets of 3 – Classical Centers/Range

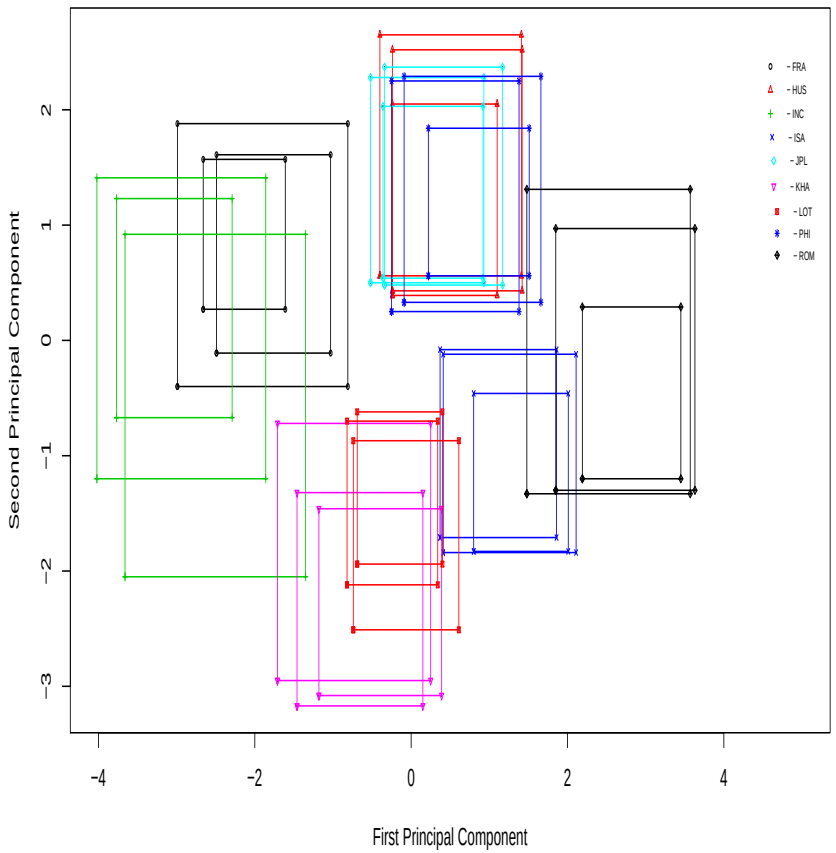


Classical on Endpoints
(Cf. Classical on Midpoints)



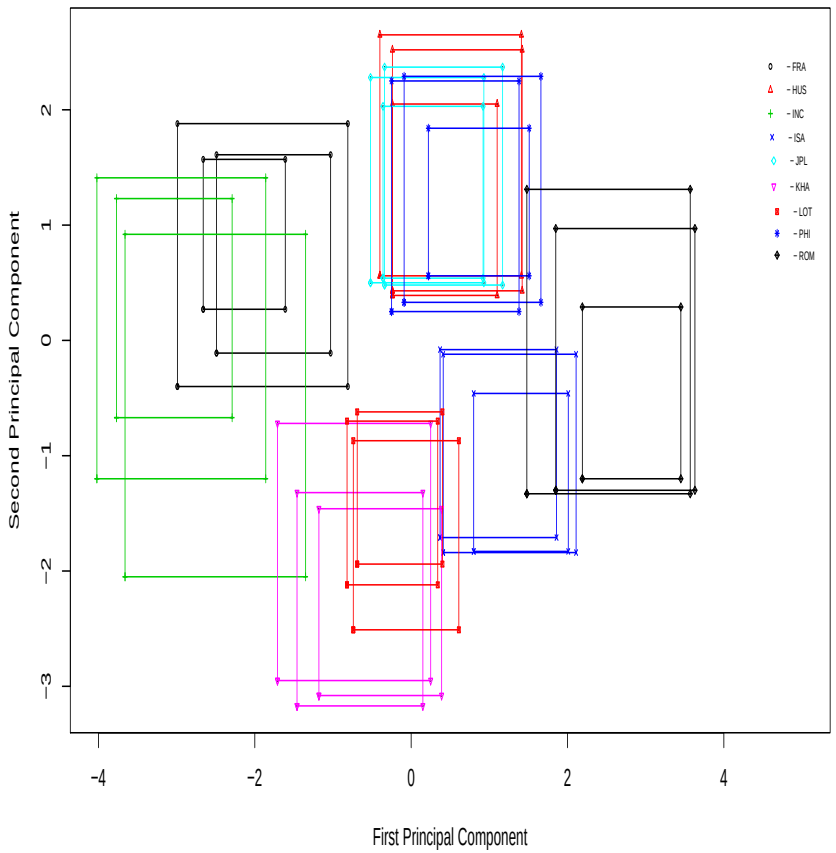
Classical on Midpoints and Range
(Some distortions, coherency lost)

Faces dataset - 27 faces in sets of 3 – Vertices and RT

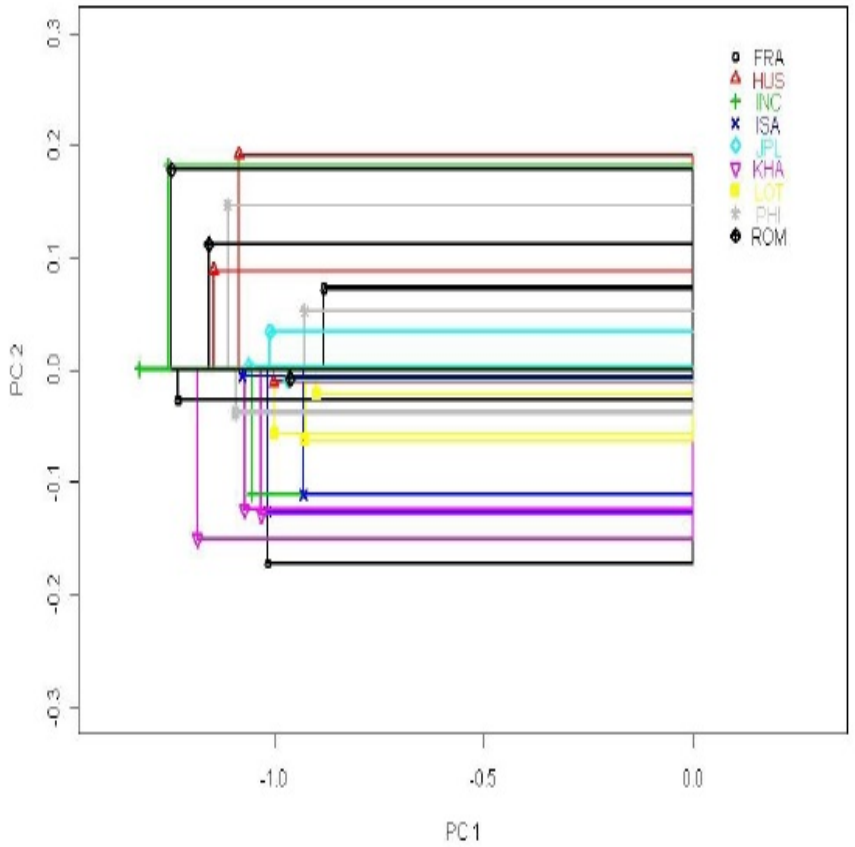


Vertices

Faces dataset - 27 faces in sets of 3 – Vertices and RT

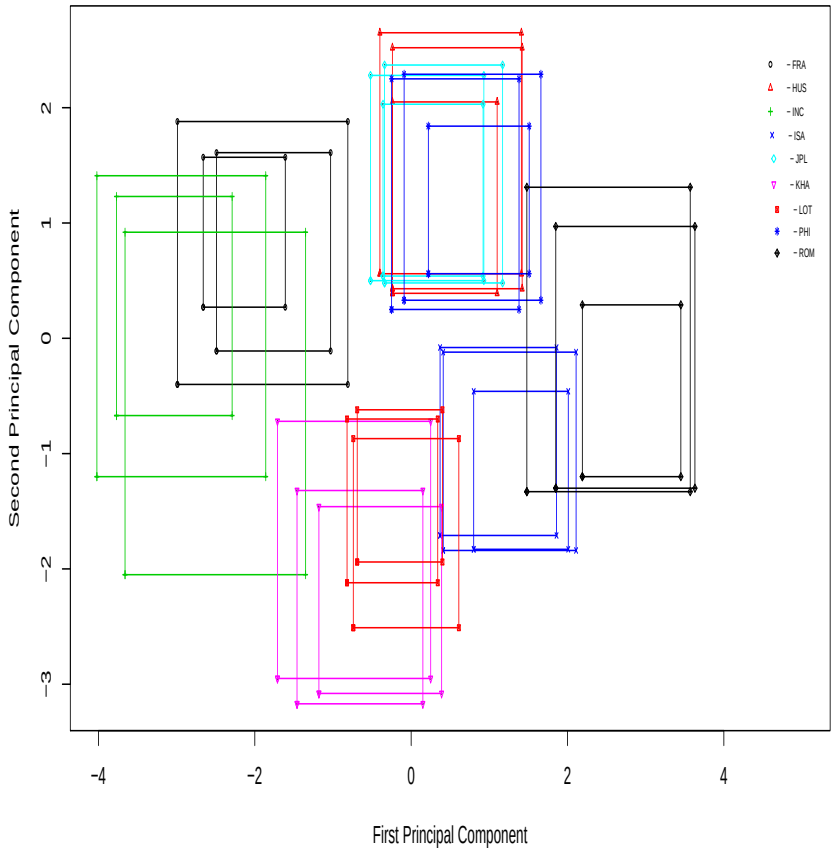


Vertices



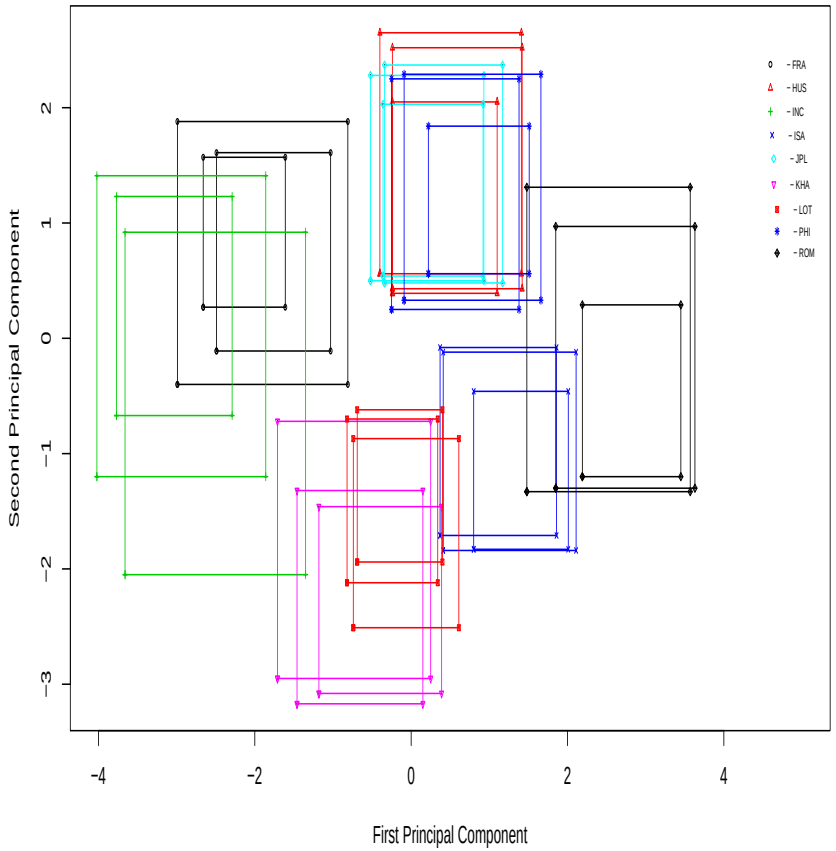
Symbolic Range-Transformation
(Distortions emerge)

Faces dataset - 27 faces in sets of 3

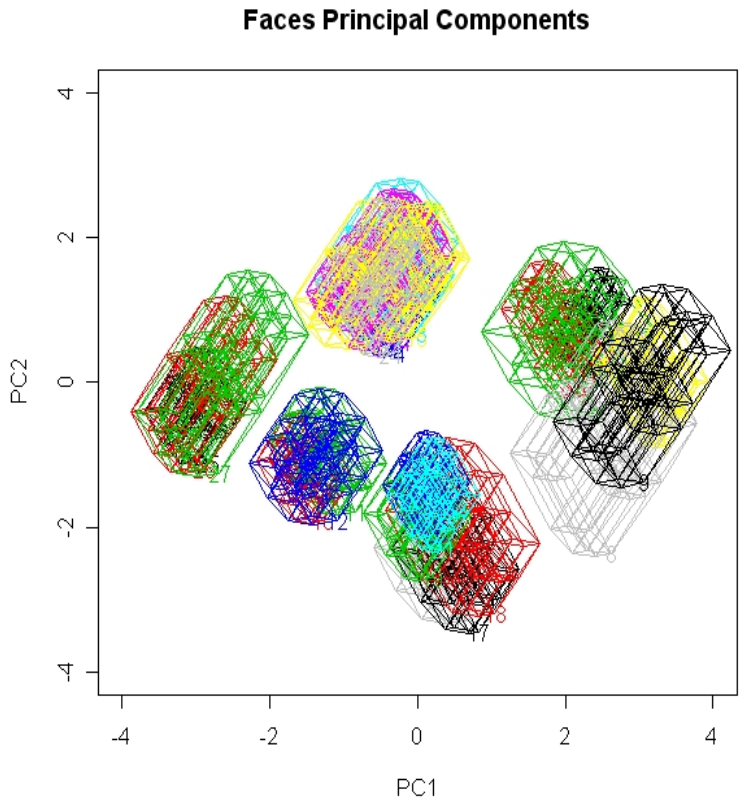


Vertices

Faces dataset - 27 faces in sets of 3



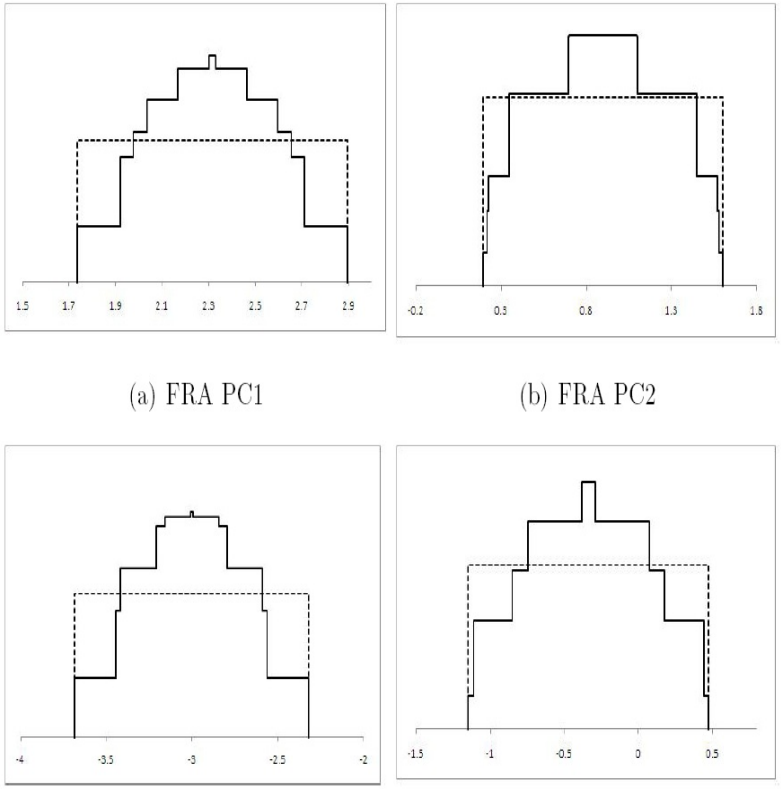
Vertices



Symbolic covariance and Polytopes

PC Output - Histograms more accurately reflect the PC output than do interval PCs

PC Output - Histograms more accurately reflect the PC output than do interval PCs

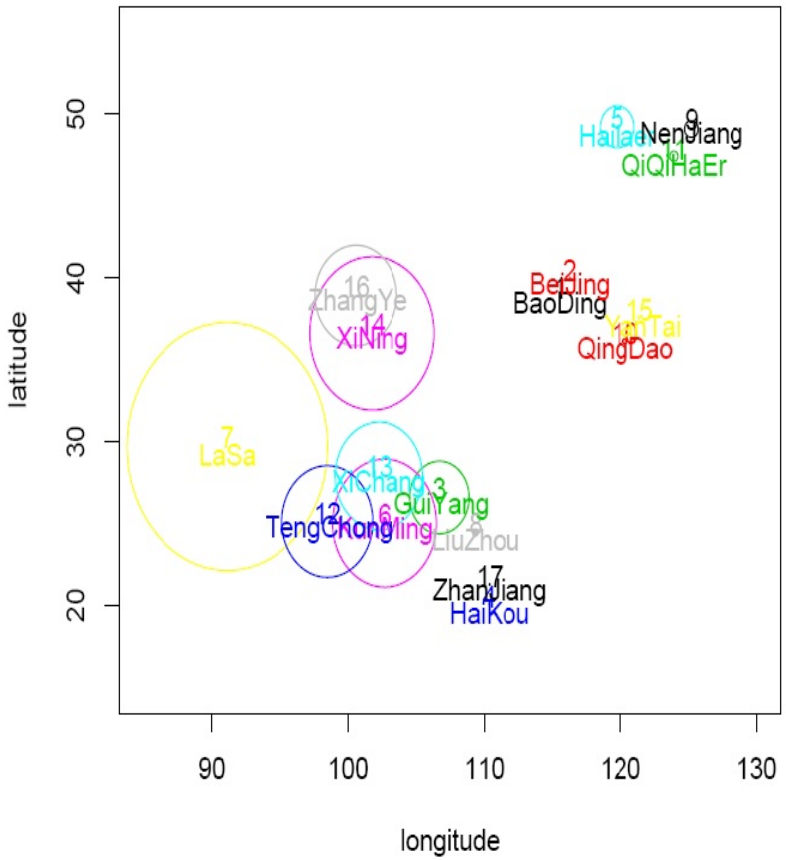


(a) FRA PC1 (b) FRA PC2

(c) ROM PC1 (d) ROM PC2

PC Output - Histograms

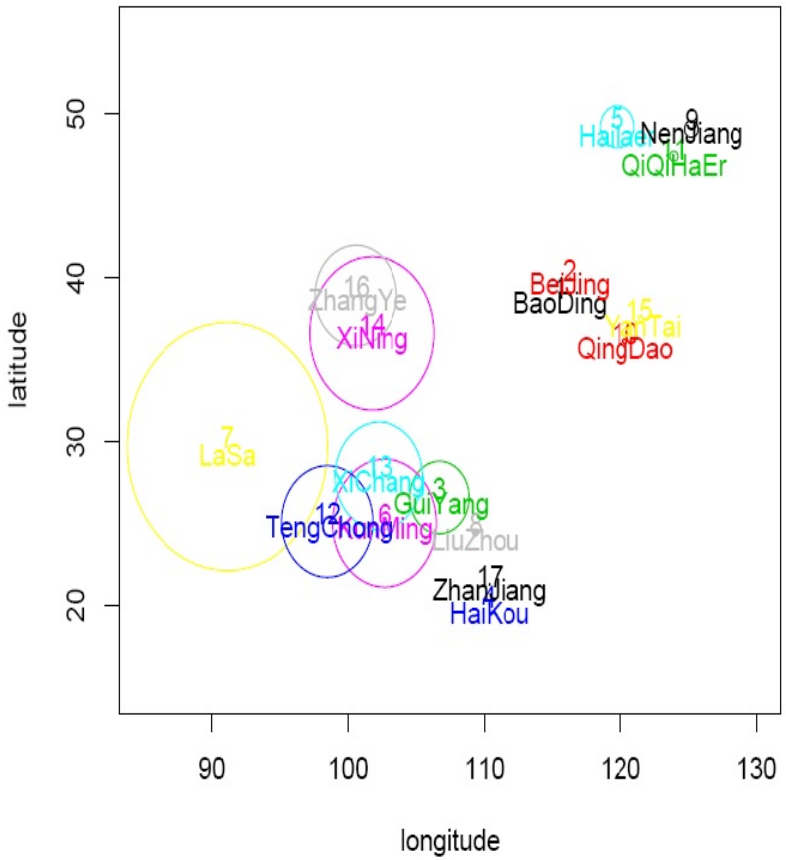
E.g. China Temperature Dataset (17 stations)



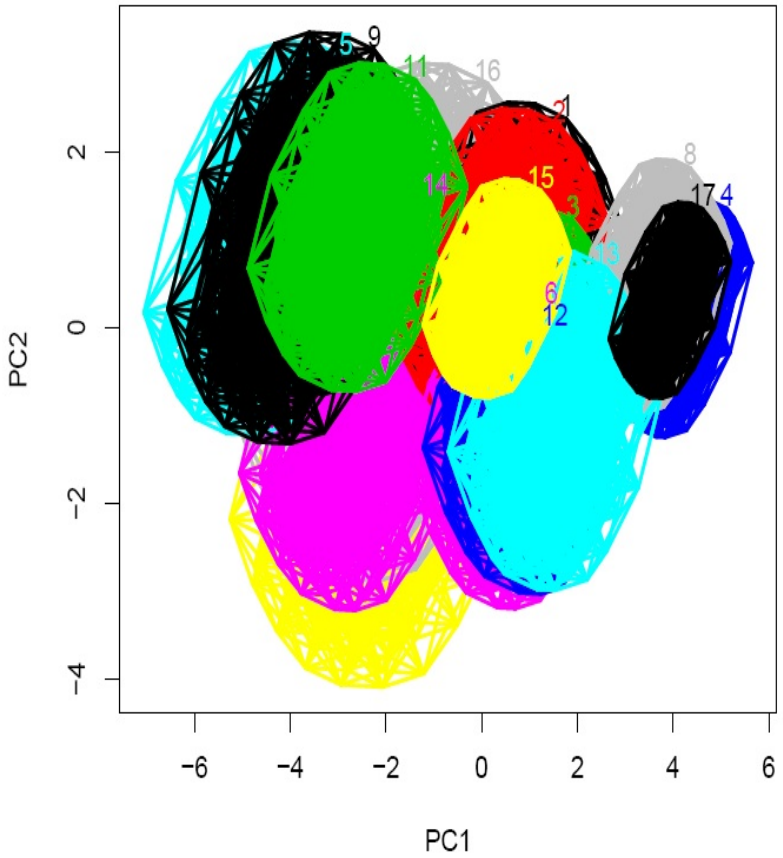
Altitude: size of location circle

PC Output - Histograms

E.g. China Temperature Dataset (17 stations)

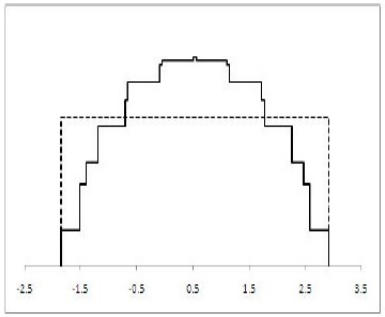


Altitude: size of location circle

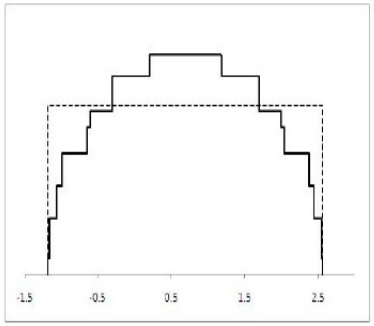


Symbolic PCA

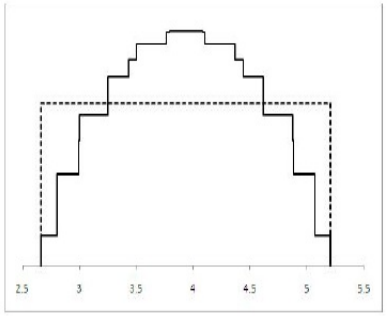
PC Output - Histograms
E.g. China Temperature Dataset (17 stations)



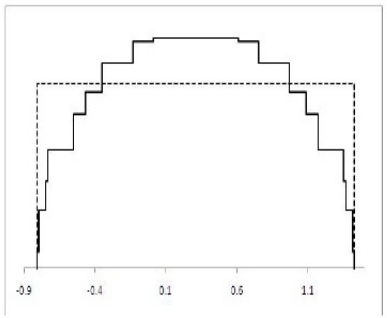
(a) BaoDing PC1



(b) BaoDing PC2



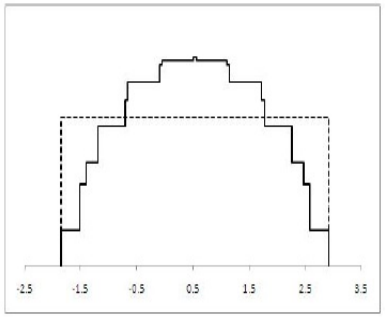
(c) ZhanJiang PC1



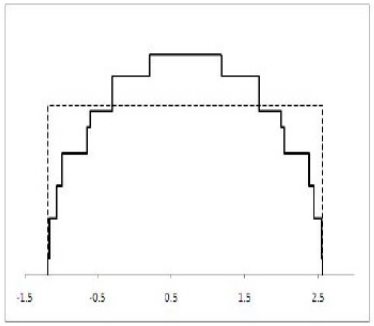
(d) ZhanJiang PC2

Le-Rademacher and Billard **(2013)**

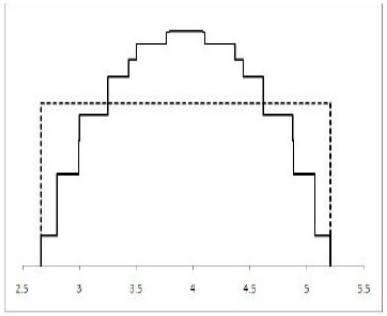
PC Output - Histograms
E.g. China Temperature Dataset (17 stations)



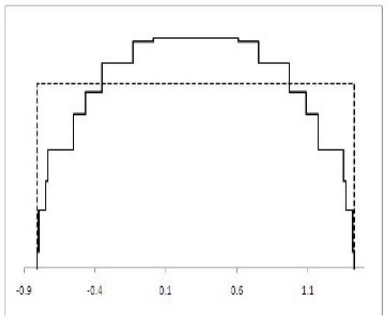
(a) BaoDing PC1



(b) BaoDing PC2



(c) ZhanJiang PC1



(d) ZhanJiang PC2

Le-Rademacher and Billard **(2013)**

Histogram data – Le-Rademacher and Billard **(2017)**

Clustering

1. Clustering for interval-valued data.
   - Partitions.
     - k-means MacQueen **(1967)**
     - k-mediods, partitioning around mediods (PAM) Kaufmann and Rousseeuw **(1987, 1990)**
     - Dynamic partitioning Diday **(1974)**, Diday and Simon **(1976)**
   - Hierarchies
     - Divisive
     - Agglomerative, including pyramids
2. Clustering for histogram-valued data.
   - Data transformation.
   - Dis/similarity measures.
   - Monothetic algorithm for histogram-valued data.
   - Polythetic algorithm for symbolic objects.
   - Cluster validity indexes.
   - Extends Chavent **(1998, 2000)**: divisive monothetic algorithm.

Clustering

1. Clustering for interval-valued data.
   - Partitions.
     - k-means MacQueen **(1967)**
     - k-mediods, partitioning around mediods (PAM) Kaufmann and Rousseeuw **(1987, 1990)**
     - Dynamic partitioning Diday **(1974)**, Diday and Simon **(1976)** ✓
   - Hierarchies
     - Divisive
     - Agglomerative, including pyramids
2. Clustering for histogram-valued data.
   - Data transformation.
   - Dis/similarity measures.
   - Monothetic algorithm for histogram-valued data.
   - Polythetic algorithm for symbolic objects.
   - Cluster validity indexes.
   - Extends Chavent **(1998, 2000)**: divisive monothetic algorithm. ✓

Clustering

1. Clustering for interval-valued data.
   - Partitions.
     - k-means MacQueen **(1967)**
     - k-mediods, partitioning around mediods (PAM) Kaufmann and Rousseeuw **(1987, 1990)**
     - Dynamic partitioning Diday **(1974)**, Diday and Simon **(1976)** ✓
   - Hierarchies
     - Divisive
     - Agglomerative, including pyramids
2. Clustering for histogram-valued data.
   - Data transformation.
   - Dis/similarity measures.
   - Monothetic algorithm for histogram-valued data.
   - Polythetic algorithm for symbolic objects.
   - Cluster validity indexes.
   - Extends Chavent **(1998, 2000)**: divisive monothetic algorithm. ✓

Billard and Diday **(2019)**

Regression-based Clustering - Interval Data

Data: $(Y, X_1, \ldots, X_p)$, with realizations, for $i = 1, \ldots, n,\ \ j = 1, \ldots, p,$

$$y_i = [y_{ia}, y_{ib}],\ x_{ij} = [x_{ija}, x_{ijb}],\ \ y_{ia} \geq y_{ib},\ \ x_{ija} \geq x_{ijb}$$

Model: $Y = \mathbf{X}'\beta + \epsilon,\quad \beta' = (\beta_0, \beta_1, \ldots, \beta_p),\ \ \mathbf{X}' = (1, X_1, \ldots, X_p),$
and ϵ is the error interval vector

Partition: $P = (C_1, \ldots, C_K),\quad n_k$ observations in C_k. Find an optimal partition that minimizes the sum of squared residuals (SSR) given K,

$$SSR = argmin_{P;\hat{\beta}_k} \sum_{k=1}^{K} \sum_{i \in C_k} r_{ki}^2 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} r_{ki}^2,$$

$$r_{ki} = d(y_i, \hat{y}_i) = d(y_i, \mathbf{x}_i'\hat{\beta}_k)$$

Regression-based Clustering - Interval Data

Data: $(Y, X_1, \ldots, X_p)$, with realizations, for $i = 1, \ldots, n,\ \ j = 1, \ldots, p,$
$y_i = [y_{ia}, y_{ib}],\ x_{ij} = [x_{ija}, x_{ijb}],\ \ y_{ia} \geq y_{ib},\ \ x_{ija} \geq x_{ijb}$

Model: $Y = \mathbf{X}'\beta + \epsilon,\quad \beta' = (\beta_0, \beta_1, \ldots, \beta_p),\ \ \mathbf{X}' = (1, X_1, \ldots, X_p),$
and ϵ is the error interval vector

Partition: $P = (C_1, \ldots, C_K),\quad n_k$ observations in C_k. Find an optimal partition that minimizes the sum of squared residuals (SSR) given K,

$$SSR = argmin_{P;\hat{\beta}_k} \sum_{k=1}^{K} \sum_{i \in C_k} r_{ki}^2 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} r_{ki}^2,$$

$$r_{ki} = d(y_i, \hat{y}_i) = d(y_i, \mathbf{x}_i'\hat{\beta}_k)$$

1. Center distance:
   $d_C(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} |x_{1j}^c - x_{2j}^c|,\ \ x_{ij}^c = (x_{ija} + x_{ijb})/2;$
2. Hausdorff distance:
   $d_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} \max\{|x_{1ja} - x_{2ja}|, |x_{1jb} - x_{2jb}|\};$
3. City-block distance:
   $d_{CB}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} [|x_{1ja} - x_{2ja}| + |x_{1jb} - x_{2jb}|].$

Regression-based Clustering

Algorithm

(i) *Initialization*: Choose a partition $P^{(0)} = (C_1^{(0)}, ..., C_K^{(0)})$ randomly from all the possible partitions, or partition the whole data set to K clusters based on some prior knowledge.

(ii) *Representation*: For $k = 1, ..., K$, fit regressions $Y_k = \boldsymbol{X}_k' \beta_k + \epsilon$ to the observations in each of the K clusters for partition $P^{(l)} = (C_1^{(l)}, ..., C_K^{(l)})$ where $l = 0, 1, \ldots,$ denotes the l^{th} iteration.

(iii) *Allocation*: For observation y_i, $i = 1, ..., n$, calculate its distance to its prediction $\hat{y}_i$ obtained by its k^{th} regression line, $d(y_i, \boldsymbol{x}_i' \hat{\beta}_k)$, $k = 1, ..., K$, and allocate the observation to its closest line; i.e., $C_k = \{(\boldsymbol{x}, y) | d(y, \boldsymbol{x}' \hat{\beta}_k) \leq d(y, \boldsymbol{x}' \hat{\beta}_{k'}), \forall k \neq k'\}$.
The updated partition is now $P^{(l+1)} = (C_1^{(l+1)}, ..., C_K^{(l+1)})$.

(iv) *Stop*: Repeat *(ii)* and *(iii)* until the improvement of SSR is smaller than a predetermined criterion, or the number of iterations reaches a predetermined maximum number.

Simulation – Method I: Naive approach –

1. Sample $\mathbf{X}^{(c)}_{n\times p}$, i.e., $\mathbf{X}^{(c)} = (1, X^{(c)}_1, \ldots, X^{(c)}_p)$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Calculate interval means $Y^{(c)} = \boldsymbol{X}^{(c)}_{n\times p}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
3. Simulate interval ranges $Y^{(r)}$, $X^{(r)}_j$ from distribution with positive support
4. Simulated observations are: $Y = [Y^{(c)} - 0.5Y^{(r)}, Y^{(c)} + 0.5Y^{(r)}]$, $X_j = [X^{(c)}_j - 0.5X^{(r)}_j, X^{(c)}_j + 0.5X^{(r)}_j]$, $j = 1, ..., p$

Simulation – Method I: Naive approach –

1. Sample $\mathbf{X}^{(c)}_{n\times p}$, i.e., $\mathbf{X}^{(c)} = (1, X^{(c)}_1, \ldots, X^{(c)}_p)$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Calculate interval means $Y^{(c)} = \boldsymbol{X}^{(c)}_{n\times p}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
3. Simulate interval ranges $Y^{(r)}$, $X^{(r)}_j$ from distribution with positive support
4. Simulated observations are: $Y = [Y^{(c)} - 0.5Y^{(r)}, Y^{(c)} + 0.5Y^{(r)}]$, $X_j = [X^{(c)}_j - 0.5X^{(r)}_j, X^{(c)}_j + 0.5X^{(r)}_j]$, $j = 1, ..., p$

Problems:

1. Interval means of Y and $\mathbf{X}$ follow regression, instead of interval-valued variables Y and $\mathbf{X}$
2. Ranges $Y^{(r)}$ independent of $\mathbf{X}^{(r)}$ - clearly cannot be true.... since - consider, for $\mathbb{X} = \{\boldsymbol{x}_i = (x_{ij}) : x_{ija} \leq x_{ij} \leq x_{ijb}, j = 1, ..., p\}$

$$y^{(r)}_i = y_{ib} - y_{ia} = \max_{\boldsymbol{x}\in\mathbb{X}}(\boldsymbol{x}'_i\boldsymbol{\beta} + \epsilon_i) - \min_{\boldsymbol{x}\in\mathbb{X}}(\boldsymbol{x}'_i\boldsymbol{\beta} + \epsilon_i)$$
$$= x^{(r)}_1|\beta_1| + ... + x^{(r)}_p|\beta_p| + \epsilon^{(r)}_i$$

$\Rightarrow$ $Y^{(r)}$ positively correlated with $\mathbf{X}^{(r)}$.

Simulation – Method I: Naive approach –

1. Sample $\mathbf{X}^{(c)}_{n\times p}$, i.e., $\mathbf{X}^{(c)} = (1, X_1^{(c)}, \ldots, X_p^{(c)})$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Calculate interval means $Y^{(c)} = \boldsymbol{X}^{(c)}_{n\times p}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
3. Simulate interval ranges $Y^{(r)}$, $X_j^{(r)}$ from distribution with positive support
4. Simulated observations are: $Y = [Y^{(c)} - 0.5Y^{(r)}, Y^{(c)} + 0.5Y^{(r)}]$, $X_j = [X_j^{(c)} - 0.5X_j^{(r)}, X_j^{(c)} + 0.5X_j^{(r)}]$, $j = 1, ..., p$

Problems:

1. Interval means of Y and $\mathbf{X}$ follow regression, instead of interval-valued variables Y and $\mathbf{X}$
2. Ranges $Y^{(r)}$ independent of $\mathbf{X}^{(r)}$ - clearly cannot be true.... since - consider, for $\mathbb{X} = \{\boldsymbol{x}_i = (x_{ij}) : x_{ija} \leq x_{ij} \leq x_{ijb}, j = 1, ..., p\}$

$$y_i^{(r)} = y_{ib} - y_{ia} = \max_{\boldsymbol{x}\in\mathbb{X}}(\boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i) - \min_{\boldsymbol{x}\in\mathbb{X}}(\boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i)$$
$$= x_1^{(r)}|\beta_1| + ... + x_p^{(r)}|\beta_p| + \epsilon_i^{(r)}$$

$\Rightarrow$ $Y^{(r)}$ positively correlated with $\mathbf{X}^{(r)}$.

Advantages:

1. Easy to implement
2. Guarantees internal distributions are uniform because assumed this to be so.

Simulation – Method II: (Compare classical simulations)

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Calculate $\boldsymbol{x}_i'\boldsymbol{\beta} = [a_i, b_i]$, $a_i = \sum_{j:\beta_j>0} x_{ija}\beta_j + \sum_{j':\beta_{j'}<0} x_{ij'b}\beta_{j'}$, $b_i = \sum_{j:\beta_j>0} x_{ijb}\beta_j + \sum_{j':\beta_{j'}<0} x_{ij'a}\beta_{j'}$
5. Add error term: $\epsilon_i^{(c)} \overset{iid}{\sim} N(0, \sigma^2)$, $\epsilon_i^{(r)} \overset{iid}{\sim} \exp(\lambda)$.
6. Calculate response interval: $y_i = [y_{ia}, y_{ib}] = [a_i + \epsilon_{ia}, b_i + \epsilon_{ib}]$

Simulation – Method II: (Compare classical simulations)

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Calculate $\boldsymbol{x}_i'\boldsymbol{\beta} = [a_i, b_i]$, $a_i = \sum_{j:\beta_j>0} x_{ija}\beta_j + \sum_{j':\beta_{j'}<0} x_{ij'b}\beta_{j'}$, $b_i = \sum_{j:\beta_j>0} x_{ijb}\beta_j + \sum_{j':\beta_{j'}<0} x_{ij'a}\beta_{j'}$
5. Add error term: $\epsilon_i^{(c)} \overset{iid}{\sim} N(0, \sigma^2)$, $\epsilon_i^{(r)} \overset{iid}{\sim} \exp(\lambda)$.
6. Calculate response interval: $y_i = [y_{ia}, y_{ib}] = [a_i + \epsilon_{ia}, b_i + \epsilon_{ib}]$

Problems:

1. Have: $y_i^{(r)} = (b_i - a_i) + (\epsilon_{ib} - \epsilon_{ia}) = (\boldsymbol{x}_i'\boldsymbol{\beta})^{(r)} + \epsilon_i^{(r)} \geq (\boldsymbol{x}_i'\boldsymbol{\beta})^{(r)}$
   $\Rightarrow$ range y_i not less than range of $\boldsymbol{x}_i'\boldsymbol{\beta}$ – Not true in practice
2. Sum of uniform distributions (here, X_{ij} and ϵ_i) is not uniform

Simulation – Method II: (Compare classical simulations)

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Calculate $\boldsymbol{x}_i'\boldsymbol{\beta} = [a_i, b_i]$, $a_i = \sum_{j:\beta_j>0} x_{ija}\beta_j + \sum_{j':\beta_{j'}<0} x_{ij'b}\beta_{j'}$, $b_i = \sum_{j:\beta_j>0} x_{ijb}\beta_j + \sum_{j':\beta_{j'}<0} x_{ij'a}\beta_{j'}$
5. Add error term: $\epsilon_i^{(c)} \overset{iid}{\sim} N(0, \sigma^2)$, $\epsilon_i^{(r)} \overset{iid}{\sim} \exp(\lambda)$.
6. Calculate response interval: $y_i = [y_{ia}, y_{ib}] = [a_i + \epsilon_{ia}, b_i + \epsilon_{ib}]$

Problems:

1. Have: $y_i^{(r)} = (b_i - a_i) + (\epsilon_{ib} - \epsilon_{ia}) = (\boldsymbol{x}_i'\boldsymbol{\beta})^{(r)} + \epsilon_i^{(r)} \geq (\boldsymbol{x}_i'\boldsymbol{\beta})^{(r)}$
   $\Rightarrow$ range y_i not less than range of $\boldsymbol{x}_i'\boldsymbol{\beta}$ – Not true in practice
2. Sum of uniform distributions (here, X_{ij} and ϵ_i) is not uniform

Advantages:

1. Same approach as for simulating classical observations.

Simulation – Method III: (Aggregation of classical data)

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Assume within these intervals uniformly distributed – then Randomly draw m obs $x_{ij1}, \ldots, x_{ijm}$ from $U(x_{ija}, x_{ijb})$
5. Take $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$, and calculate $y_i = [y_{ia}, y_{ib}]$,

$$y_{ia} = \min_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\},$$
$$y_{ib} = \max_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\}$$

Here, m predetermined, or better yet $m_i \overset{iid}{\sim} f(m; \lambda)$

Simulation – Method III: (Aggregation of classical data)

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Assume within these intervals uniformly distributed – then Randomly draw m obs $x_{ij1}, \ldots, x_{ijm}$ from $U(x_{ija}, x_{ijb})$
5. Take $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$, and calculate $y_i = [y_{ia}, y_{ib}]$,

$$y_{ia} = \min_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\},$$
$$y_{ib} = \max_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\}$$

Here, m predetermined, or better yet $m_i \overset{iid}{\sim} f(m; \lambda)$

Problem:

Cannot guarantee uniformity within y_i intervals

Simulation – Method III: (Aggregation of classical data)

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Assume within these intervals uniformly distributed – then Randomly draw m obs $x_{ij1}, \ldots, x_{ijm}$ from $U(x_{ija}, x_{ijb})$
5. Take $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$, and calculate $y_i = [y_{ia}, y_{ib}]$,

$$y_{ia} = \min_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\},$$
$$y_{ib} = \max_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\}$$

Here, m predetermined, or better yet $m_i \overset{iid}{\sim} f(m; \lambda)$

Problem:

Cannot guarantee uniformity within y_i intervals

Advantage:

Close to how intervals collected in practice.

Simulation – Method IV: (Aggregation of classical data)

1-5. Same as Method III, except that take m large, eg $m \geq 3000$ (Xu **(2014)**, Cariou and B **(2015)**), i.e.,

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Assume within these intervals uniformly distributed – then Randomly draw m obs $x_{ij1}, \ldots, x_{ijm}$ from $U(x_{ija}, x_{ijb})$
5. Take $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$, and calculate $y_i = [y_{ia}, y_{ib}]$,
$$y_{ia} = \min_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\},$$
$$y_{ib} = \max_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\}$$
6. Take y_i as interval from 1st to 3rd quantile

Simulation – Method IV: (Aggregation of classical data)

1-5. Same as Method III, except that take m large, eg $m \geq 3000$ (Xu **(2014)**, Cariou and B **(2015)**), i.e.,

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Assume within these intervals uniformly distributed – then Randomly draw m obs $x_{ij1}, \ldots, x_{ijm}$ from $U(x_{ija}, x_{ijb})$
5. Take $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$, and calculate $y_i = [y_{ia}, y_{ib}]$,
$$y_{ia} = \min_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\},$$
$$y_{ib} = \max_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\}$$
6. Take y_i as interval from 1st to 3rd quantile

Problem:

m not always large enough to ensure uniformity

Simulation – Method IV: (Aggregation of classical data)

1-5. Same as Method III, except that take m large, eg $m \geq 3000$ (Xu **(2014)**, Cariou and B **(2015)**), i.e.,

1. Sample predictor means $\mathbf{X}^{(c)}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. Sample predictor ranges $\mathbf{X}^{(r)}$ from distributions with positive support
3. This gives predictor variables $x_{ij} = [x_{ij}^{(c)} - 0.5x_{ij}^{(r)}, x_{ij}^{(c)} + 0.5x_{ij}^{(r)}]$
4. Assume within these intervals uniformly distributed – then Randomly draw m obs $x_{ij1}, \ldots, x_{ijm}$ from $U(x_{ija}, x_{ijb})$
5. Take $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$, and calculate $y_i = [y_{ia}, y_{ib}]$,
$$y_{ia} = \min_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\},$$
$$y_{ib} = \max_{l \in \{1,\ldots,m\}} \{\beta_0 + \beta_1 x_{i1l} + \ldots + \beta_p x_{ipl} + \epsilon_{il}\}$$
6. Take y_i as interval from 1st to 3rd quantile

Problem:

m not always large enough to ensure uniformity

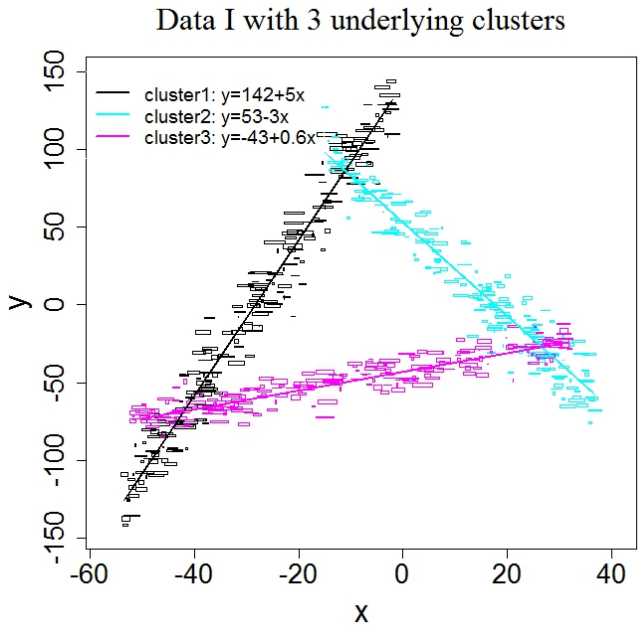
Advantage:

Removes problem of Method III. Close to how intervals collected in practice.

k-means v k-regressions methods

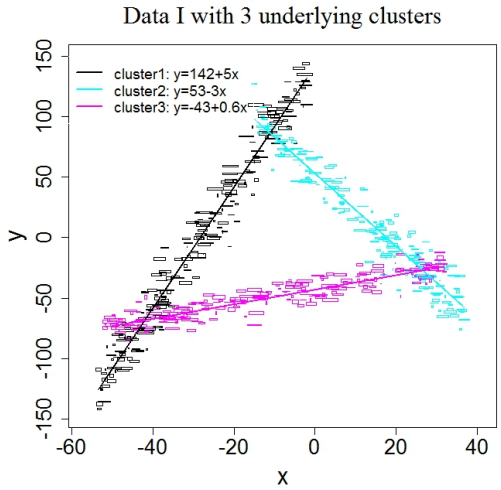
Data set (I) is composed of three clusters that follow the equations:

$$(1) : y = 142 + 5x + \epsilon_1,$$
$$(2) : y = 53 - 3x + \epsilon_2,$$
$$(3) : y = -43 + 0.6x + \epsilon_3$$

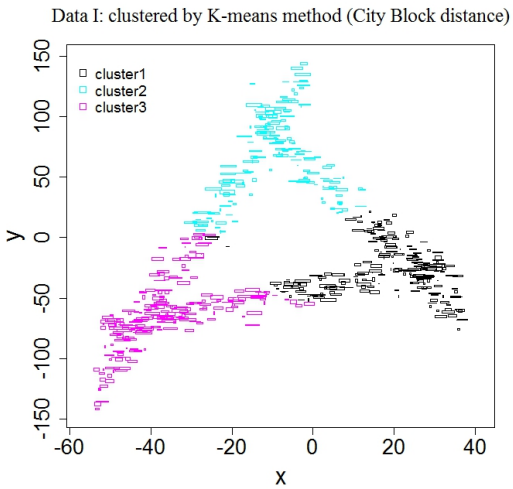


(Method III, $m = 25$, $\epsilon_1 \sim N(0, 15^2)$, $\epsilon_2 \sim N(0, 12^2)$, $\epsilon_3 \sim N(0, 7^2)$)

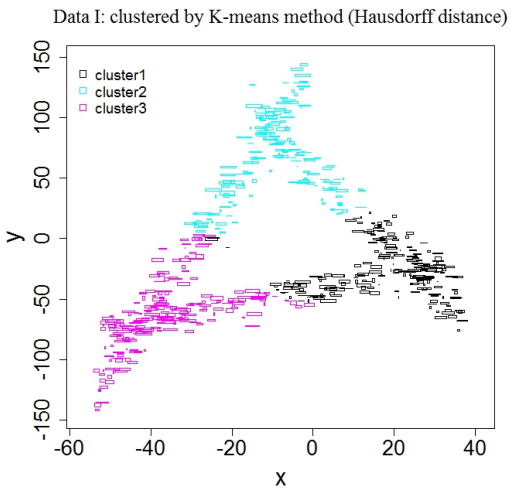
k-means algorithm:



Original Data

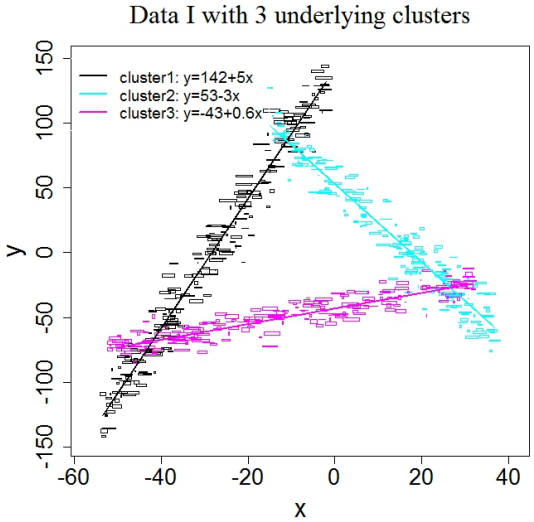


City block

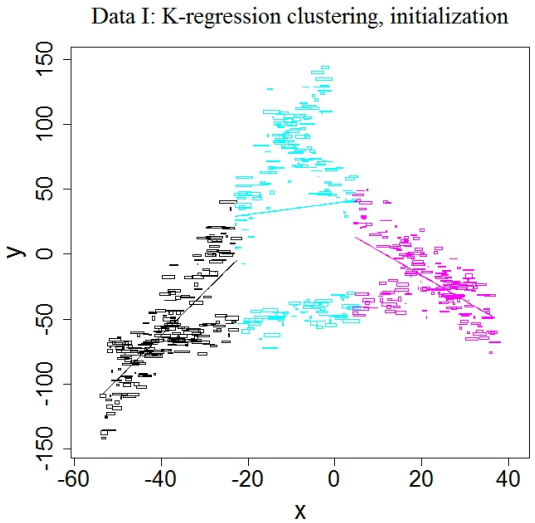


Hausdorff

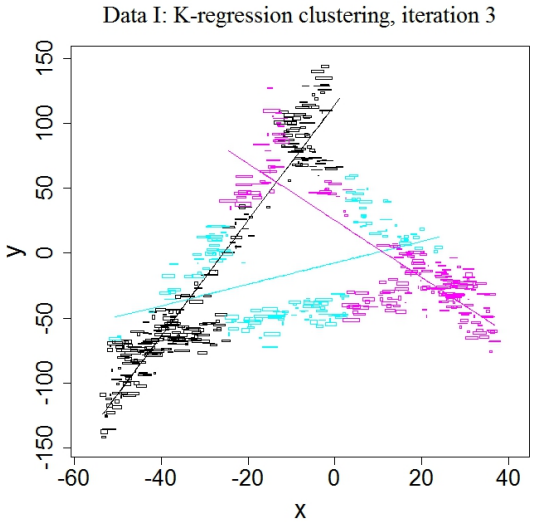
k-regressions algorithm:



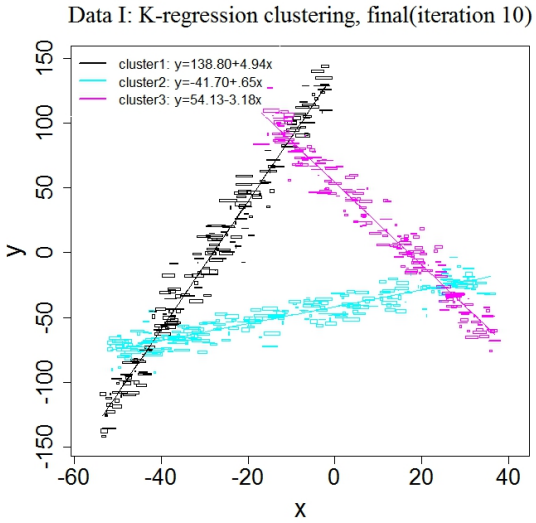
Original Data



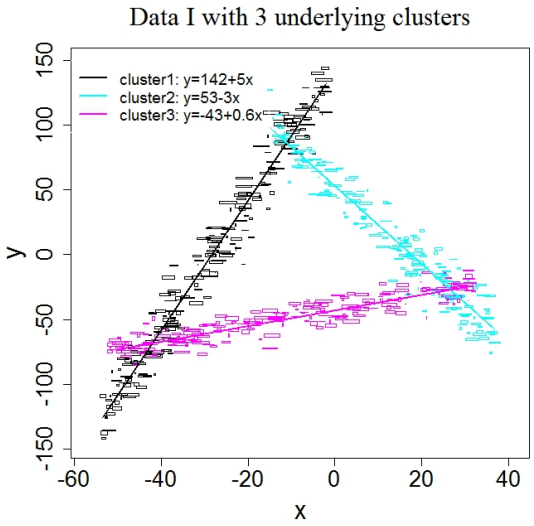
Initialization



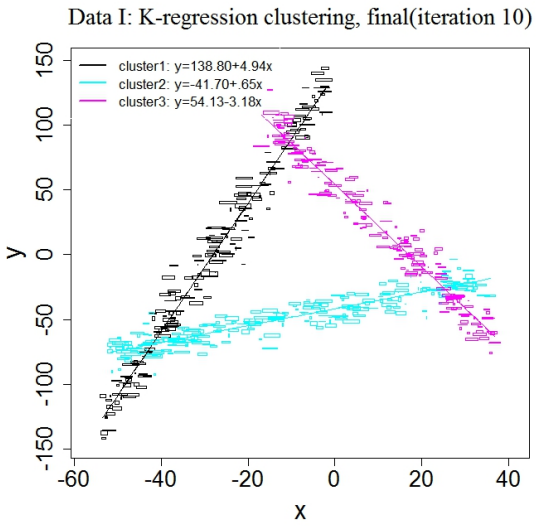
iteration 3



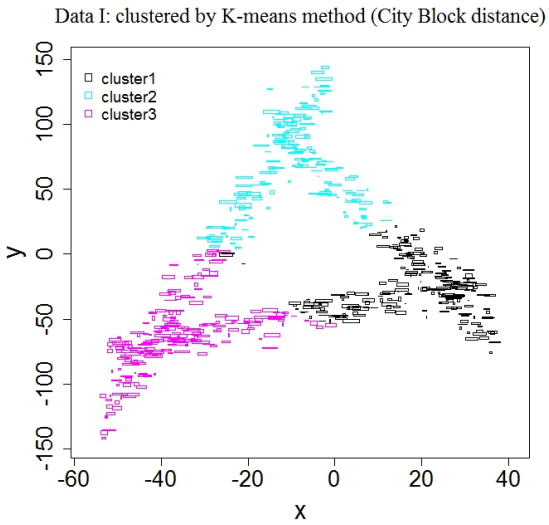
Final - iteration 10



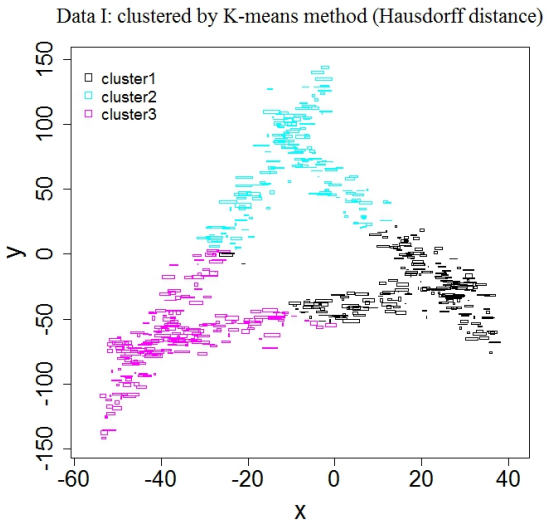
Original Data



k-regressions (center)



k-means (city block)

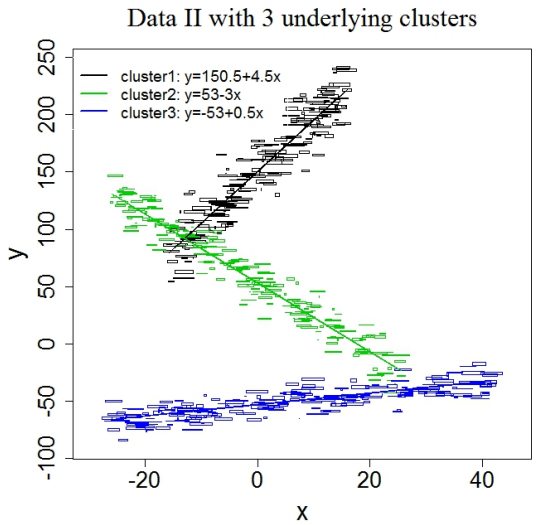


k-means (Hausdorff)

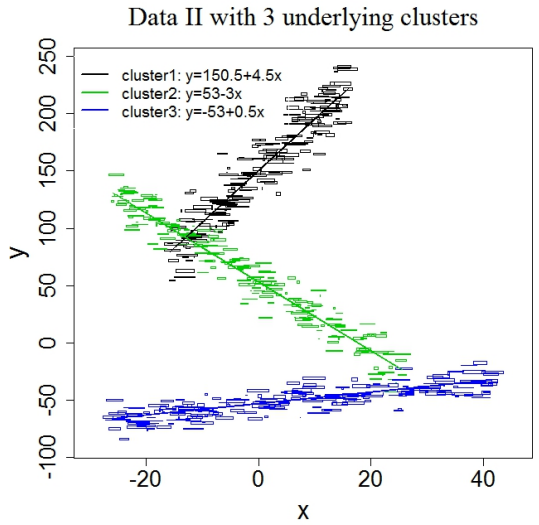
k-means v k-regressions methods

Data set (II) is composed of three clusters that follow the equations:

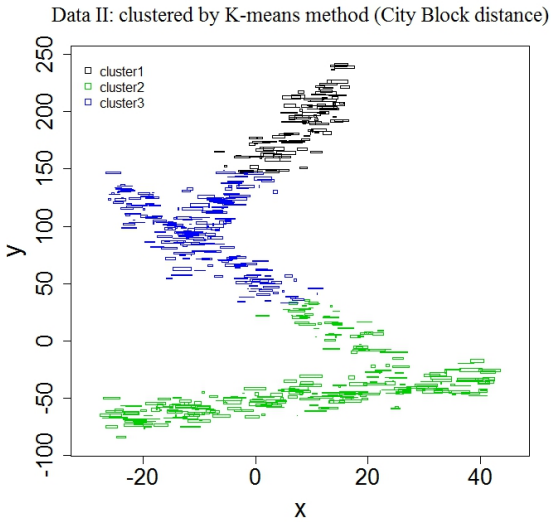
$$(1) : y = 150.5 + 4.5x + \epsilon_1,$$
$$(2) : y = 53 - 3x + \epsilon_2,$$
$$(3) : y = -53 + 0.5x + \epsilon_3$$



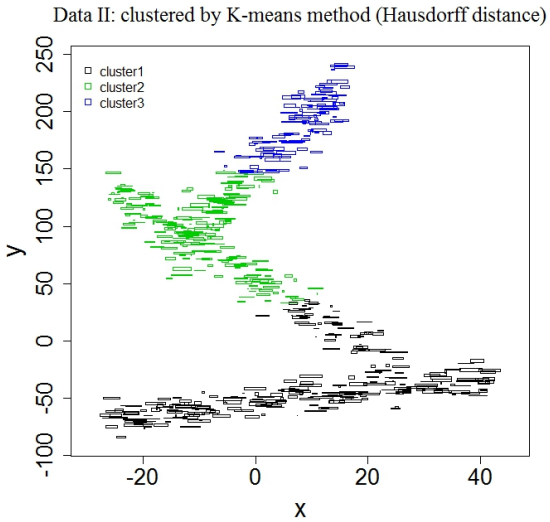
(Method IV, $m = 3000$, $\epsilon_1 \sim N(0, 15^2)$, $\epsilon_2 \sim N(0, 12^2)$, $\epsilon_3 \sim N(0, 7^2)$)



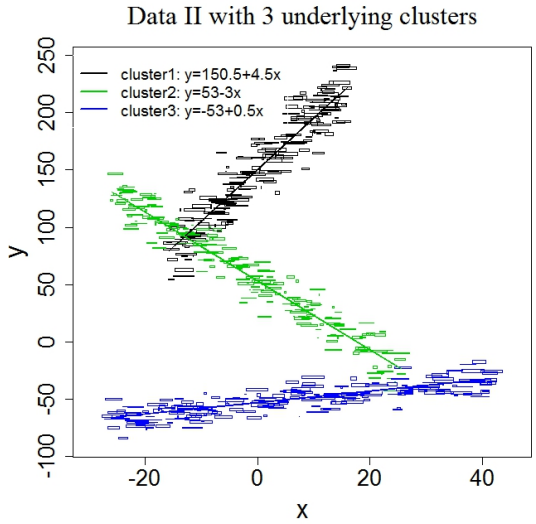
Original data



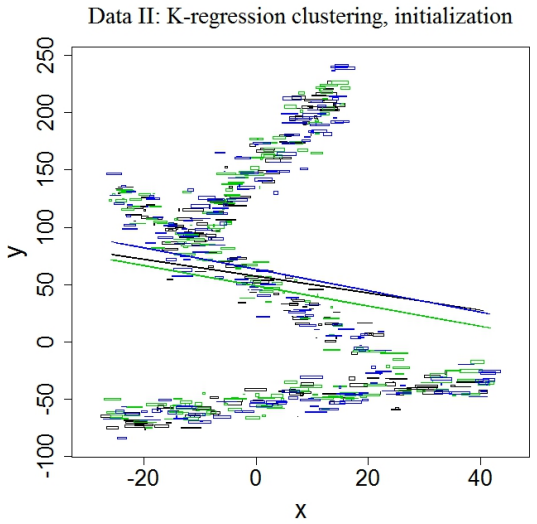
k-means (city block)



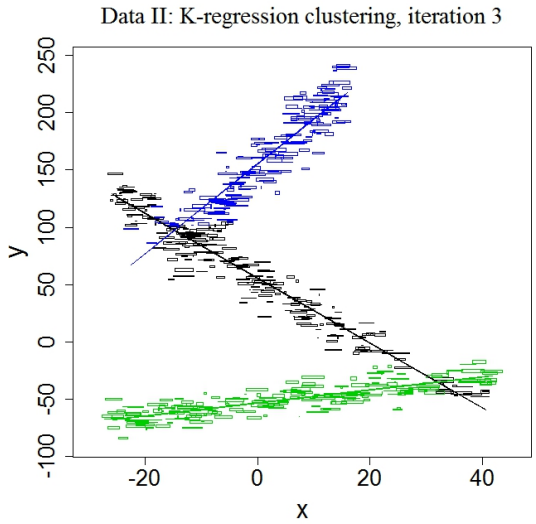
k-means (Hausdorff)



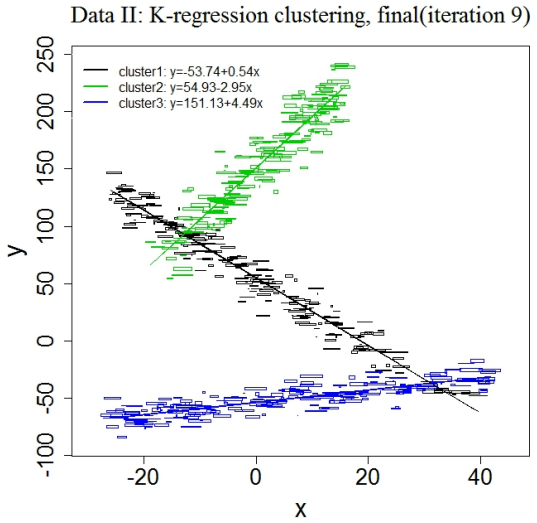
Original data



Initialization



Iteration 3



Final - iteration 9

Three different data structures:

(1) $y = 1.0 + 1.3x$
(2) $y = 45 + 1.8x$
(3) $y = 45 - 2.5x$

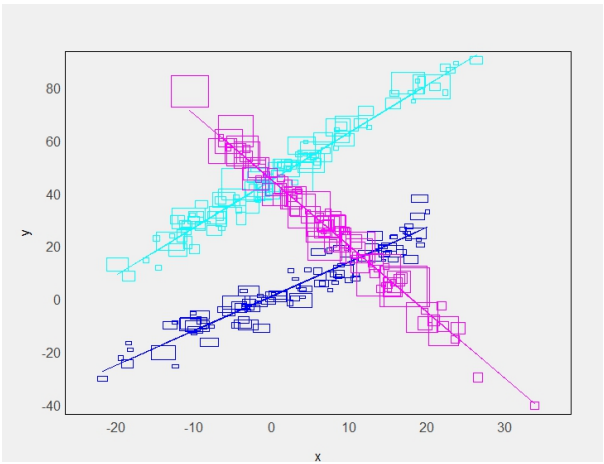
Data A

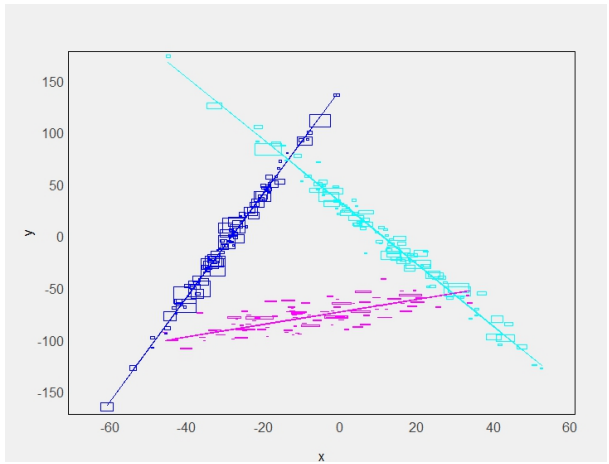
(1) $y = 142 + 5x$
(2) $y = 33 - 3x$
(3) $y = -73 + 0.6x$

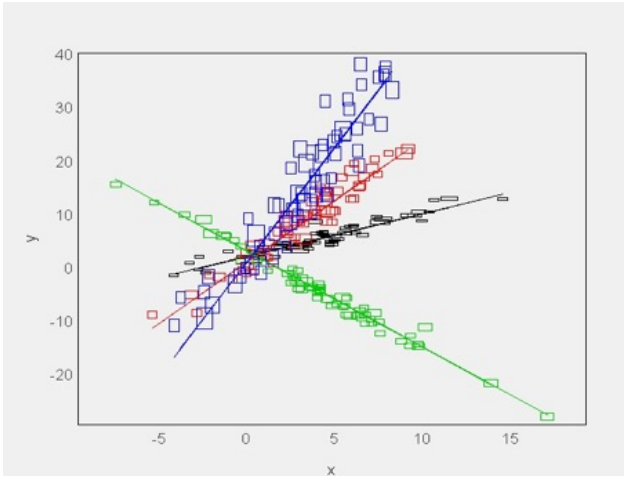
Data B

(1) $y = 2.0 + 0.8x$
(2) $y = 1.0 + 2.3x$
(3) $y = 3.0 - 1.8x$
(4) $y = 1.0 + 4.3x$

Data C







k-regressions clustering results for Data *A* (# replications=100)

		True	center		city-block		Hausdorff	
	β_j	Values	mean	std	mean	std	mean	std
Cluster 1	β_0	1.00	0.92	0.55	0.84	0.55	0.91	0.64
	β_1	1.30	1.31	0.04	1.29	0.05	1.29	0.05
Cluster 2	β_0	45.00	45.02	0.46	45.02	0.45	44.90	0.46
	β_1	1.80	1.80	0.04	1.80	0.04	1.80	0.04
Cluster 3	β_0	45.00	44.86	0.50	45.13	0.53	45.12	0.50
	β_1	-2.50	-2.49	0.04	-2.50	0.04	-2.51	0.04
SSR	-	-	890.26	39.18	2181.25	86.35	1458.43	50.66

k-regressions clustering results for Data B (# replications=100)

		True	center		city-block		Hausdorff	
	β_j	Values	mean	std	mean	std	mean	std
Cluster 1	β_0	142.00	141.30	2.09	140.77	1.89	141.75	2.01
	β_1	5.00	4.97	0.07	4.95	0.07	4.99	0.07
Cluster 2	β_0	33.00	33.06	1.25	33.39	1.30	33.42	1.28
	β_1	-3.00	-2.99	0.06	-3.00	0.06	-2.99	0.06
Cluster 3	β_0	-73.00	-72.93	0.98	-72.83	1.11	-72.64	1.14
	β_1	0.60	0.60	0.05	0.60	0.05	0.60	0.06
SSR	-	-	1757.41	85.17	4127.63	177.30	2689.33	90.65

k-regressions clustering results for Data *C* (# replications=100)

		True	center		city-block		Hausdorff	
	β_j	Values	mean	std	mean	std	mean	std
Cluster 1	β_0	2.00	2.06	0.38	3.55	1.76	3.86	2.98
	β_1	0.80	0.81	0.05	0.68	0.17	0.73	0.18
Cluster 2	β_0	1.00	1.32	1.31	3.06	2.74	5.20	4.04
	β_1	2.30	2.36	0.22	2.28	0.50	1.97	0.73
Cluster 3	β_0	3.00	2.90	0.32	3.12	0.34	3.02	0.39
	β_1	-1.80	-1.78	0.04	-1.81	0.05	-1.80	0.05
Cluster 4	β_0	1.00	2.13	1.87	4.29	2.67	4.24	2.82
	β_1	4.30	4.27	0.35	4.04	0.46	4.05	0.48
SSR	-	-	296.25	20.52	777.90	44.84	521.13	27.75

Real data: Faces data of Leroy et al. **(1996)**

Y = length nose-bridge, X = eye span

Cluster 1 (black):

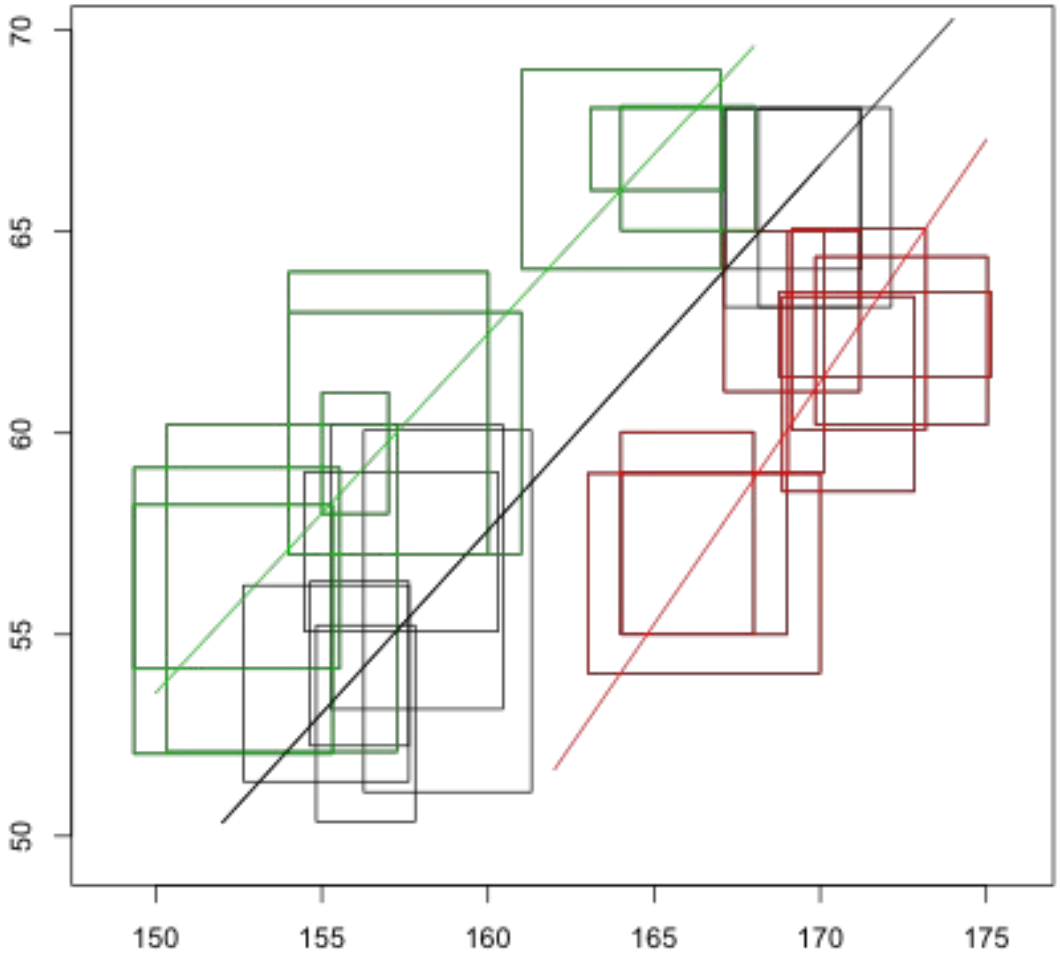
$Y = -87.52 + 0.91X + \epsilon$

Cluster 2 (red):

$Y = -143.19 + 1.20X + \epsilon$

Cluster 3 (green):

$Y = -80.17 + 0.89X + \epsilon$



k-regressions clusters for the faces data with $K = 3$

Clustering

1. Clustering for interval-valued data.
   - Partitions.
     - k-means MacQueen **(1967)**
     - k-mediods, partitioning around mediods (PAM) Kaufmann and Rousseeuw **(1987, 1990)**
     - Dynamic partitioning Diday **(1974)**, Diday and Simon **(1976)** ✓
   - Hierarchies
     - Divisive
     - Agglomerative, including pyramids
2. Clustering for histogram-valued data.
   - Data transformation.
   - Dis/similarity measures.
   - Monothetic algorithm for histogram-valued data.
   - Polythetic algorithm for symbolic objects.
   - Cluster validity indexes.
   - Extends Chavent **(1998, 2000)**: divisive monothetic algorithm. ✓

Billard and Diday **(2019)**

Clustering

1. Clustering for interval-valued data. ✓
   - Partitions.
     - k-means MacQueen **(1967)**
     - k-mediods, partitioning around mediods (PAM) Kaufmann and Rousseeuw **(1987, 1990)**
     - Dynamic partitioning Diday **(1974)**, Diday and Simon **(1976)** ✓
   - Hierarchies
     - Divisive
     - Agglomerative, including pyramids
2. Clustering for histogram-valued data.
   - Data transformation.
   - Dis/similarity measures.
   - Monothetic algorithm for histogram-valued data.
   - Polythetic algorithm for symbolic objects.
   - Cluster validity indexes.
   - Extends Chavent **(1998, 2000)**: divisive monothetic algorithm. ✓

Billard and Diday **(2019)**

Histogram Observations

Data:

$$y_i = \left\{ [a_{ijk}, a_{ij,k+1}),\ p_{ijk};\ j = 1, \ldots, p,\ k = 1, \ldots, v_{ij} \right\},\ \sum_{k=1}^{v_{ij}} p_{ijk} = 1.$$

Histogram Observations

Data:

$$y_i = \left\{ [a_{ijk}, a_{ij,k+1}),\ p_{ijk};\ j = 1, \ldots, p,\ k = 1, \ldots, v_{ij} \right\},\ \sum_{k=1}^{v_{ij}} p_{ijk} = 1.$$

Transformed Data:

Let $\{[b_{jk}, b_{j,k+1}),\ k = 1, 2, \ldots, t_j\}$ be transformed subintervals for Y_j. Then

$$b_{j1} = \min_i \{a_{ij1}\},\quad b_{j,t_j+1} = \max_i \{a_{ij,t_{ij}+1}\},$$

$$b_{j,k+1} = b_{j1} + k\Psi_j/t_j,\ k = 1, \ldots, t_j,$$

where

$$\Psi_j = b_{j,t_j+1} - b_{j1}$$

and

$$t_j = \left\lceil \frac{\Psi_j}{\min_{i,k}\{a_{ij,k+1} - a_{ijk}\}} \right\rfloor,$$

where $\lceil \cdot \rfloor$ is rounding off a number to the nearest integer. Thus, a transformed histogram-valued observation is

$$y_i' = \{[b_{jk}, b_{j,k+1}),\ p_{ijk}';\ j = 1, \ldots, p,\ k = 1, \ldots, t_j\},\ i = 1, \ldots, n,\ \sum_{k=1}^{t_j} p_{ijk}' = 1$$

Hierarchical Divisive Clustering Methods - Some Notation

- $\Omega \ni y_i = \{y_{i1}, \ldots, y_{ip}\},\ i = 1, \ldots, n.$
- $C_u = \{y_1, \ldots, y_{n_u}\}.$
- P_r : a partition of Ω at the r^{th} stage.
- $P_r = \{C_u,\ u = 1, \ldots, r\}.$
- $P_{r+1} = \left(P_r \cup \{C_u^1, C_u^2\}\right) - \{C_u\}.$

Hierarchical Divisive Clustering Methods - Some Notation

- $\Omega \ni y_i = \{y_{i1}, \ldots, y_{ip}\},\ i = 1, \ldots, n.$
- $C_u = \{y_1, \ldots, y_{n_u}\}.$
- P_r : a partition of Ω at the r^{th} stage.
- $P_r = \{C_u,\ u = 1, \ldots, r\}.$
- $P_{r+1} = \left(P_r \cup \{C_u^1, C_u^2\}\right) - \{C_u\}.$

Double algorithm:

1. partition by means $\rightsquigarrow C_u = (C_u^{M_j,1}, C_u^{M_j,2})$
2. partition by variances $\rightsquigarrow C_u = (C_u^{S_j,1}, C_u^{S_j,2})$

Within-Cluster and Between-Cluster Variance

Within-cluster variance, with weight w_i for y_i,

$$I(C_u) = \frac{1}{2\tau} \sum_{i_1=1}^{n_u} \sum_{i_2=1}^{n_u} w_{i_1} w_{i_2} D^2(y_{i_1}, y_{i_2}), \quad \tau = \sum_{i=1}^{n_u} w_i.$$

Total within-cluster variance : $W(P_r) = \sum_{u=1}^{r} I(C_u)$.
Between-cluster variance : $B(P_r) = W(\Omega) - W(P_r)$.

Within-Cluster and Between-Cluster Variance

Within-cluster variance, with weight w_i for y_i,

$$I(C_u) = \frac{1}{2\tau} \sum_{i_1=1}^{n_u} \sum_{i_2=1}^{n_u} w_{i_1} w_{i_2} D^2(y_{i_1}, y_{i_2}), \quad \tau = \sum_{i=1}^{n_u} w_i.$$

Total within-cluster variance : $W(P_r) = \sum_{u=1}^{r} I(C_u)$.
Between-cluster variance : $B(P_r) = W(\Omega) - W(P_r)$.

Double algorithm: $C_u = (C_{u,1}, \ldots, C_{u,q}, C_{u,q+1}, \ldots, C_{u,n_u})$

1. partition by means: $\Delta_{u,q}^{M_j} = I(C_u) - I(C_{u,q}^{M_j,1}) - I(C_{u,q}^{M_j,2})$
2. partition by variances: $\Delta_{u,q}^{S_j} = I(C_u) - I(C_{u,q}^{S_j,1}) - I(C_{u,q}^{S_j,2})$

Within-Cluster and Between-Cluster Variance

Within-cluster variance, with weight w_i for y_i,

$$I(C_u) = \frac{1}{2\tau} \sum_{i_1=1}^{n_u} \sum_{i_2=1}^{n_u} w_{i_1} w_{i_2} D^2(y_{i_1}, y_{i_2}), \quad \tau = \sum_{i=1}^{n_u} w_i.$$

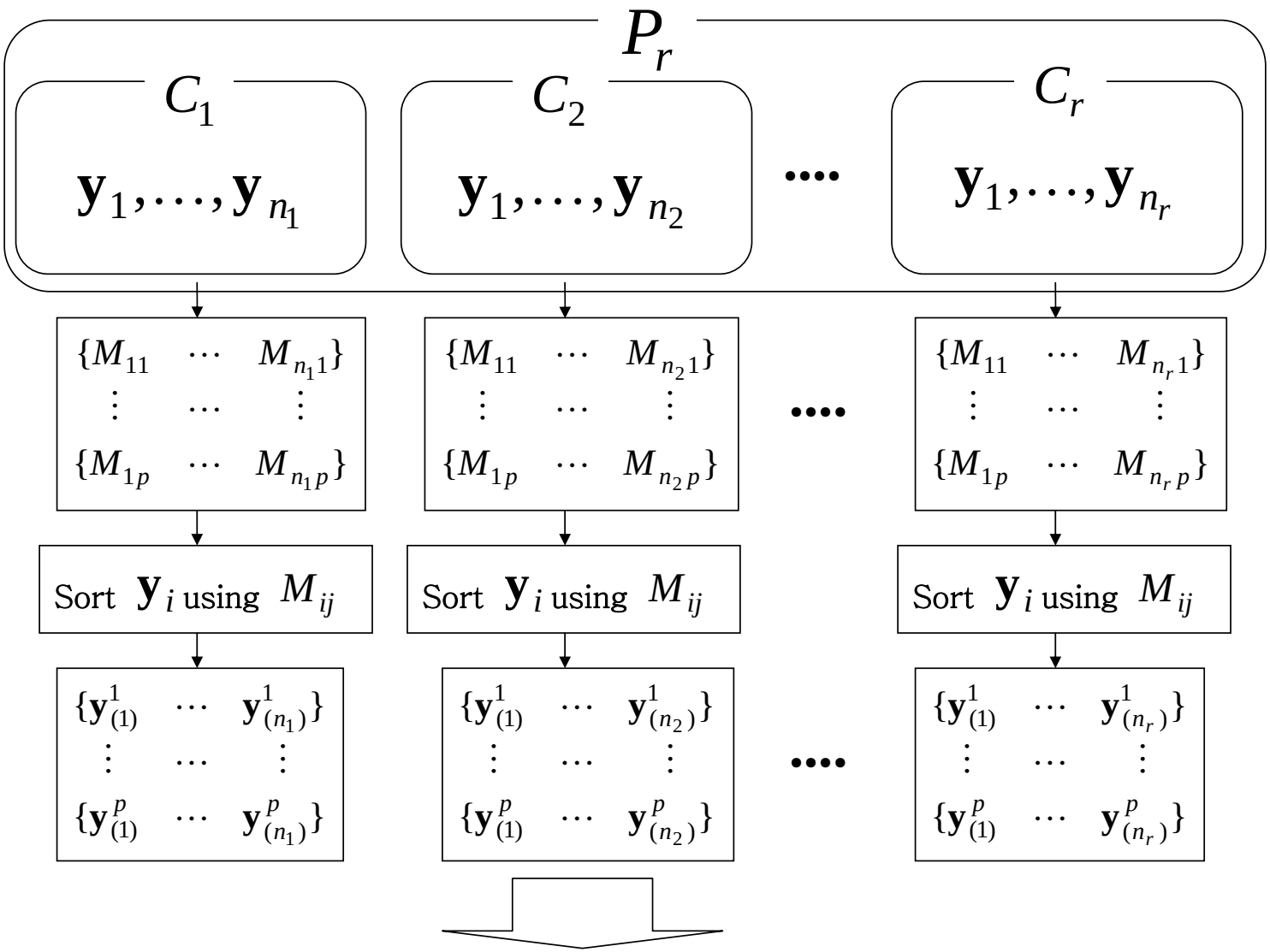
Total within-cluster variance : $W(P_r) = \sum_{u=1}^{r} I(C_u)$.
Between-cluster variance : $B(P_r) = W(\Omega) - W(P_r)$.

Double algorithm: $C_u = (C_{u,1}, \ldots, C_{u,q}, C_{u,q+1}, \ldots, C_{u,n_u})$

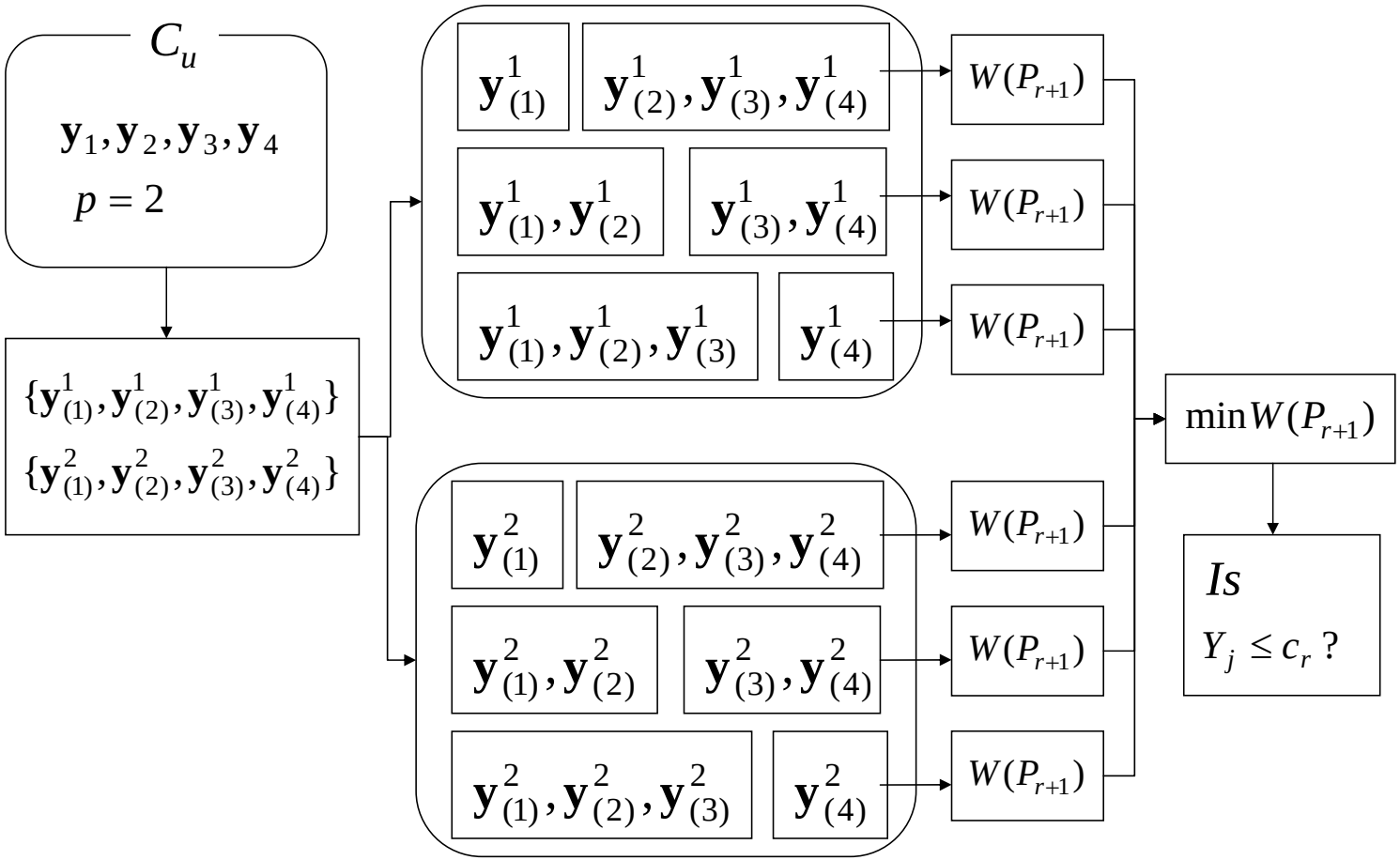
1. partition by means: $\Delta_{u,q}^{M_j} = I(C_u) - I(C_{u,q}^{M_j,1}) - I(C_{u,q}^{M_j,2})$
2. partition by variances: $\Delta_{u,q}^{S_j} = I(C_u) - I(C_{u,q}^{S_j,1}) - I(C_{u,q}^{S_j,2})$

Select: $\Delta = \max\{\Delta_{u,q}^{M_j}, \Delta_{u,q}^{S_j}, \ u, q, j\}$

Monothetic Algorithm for Symbolic Objects

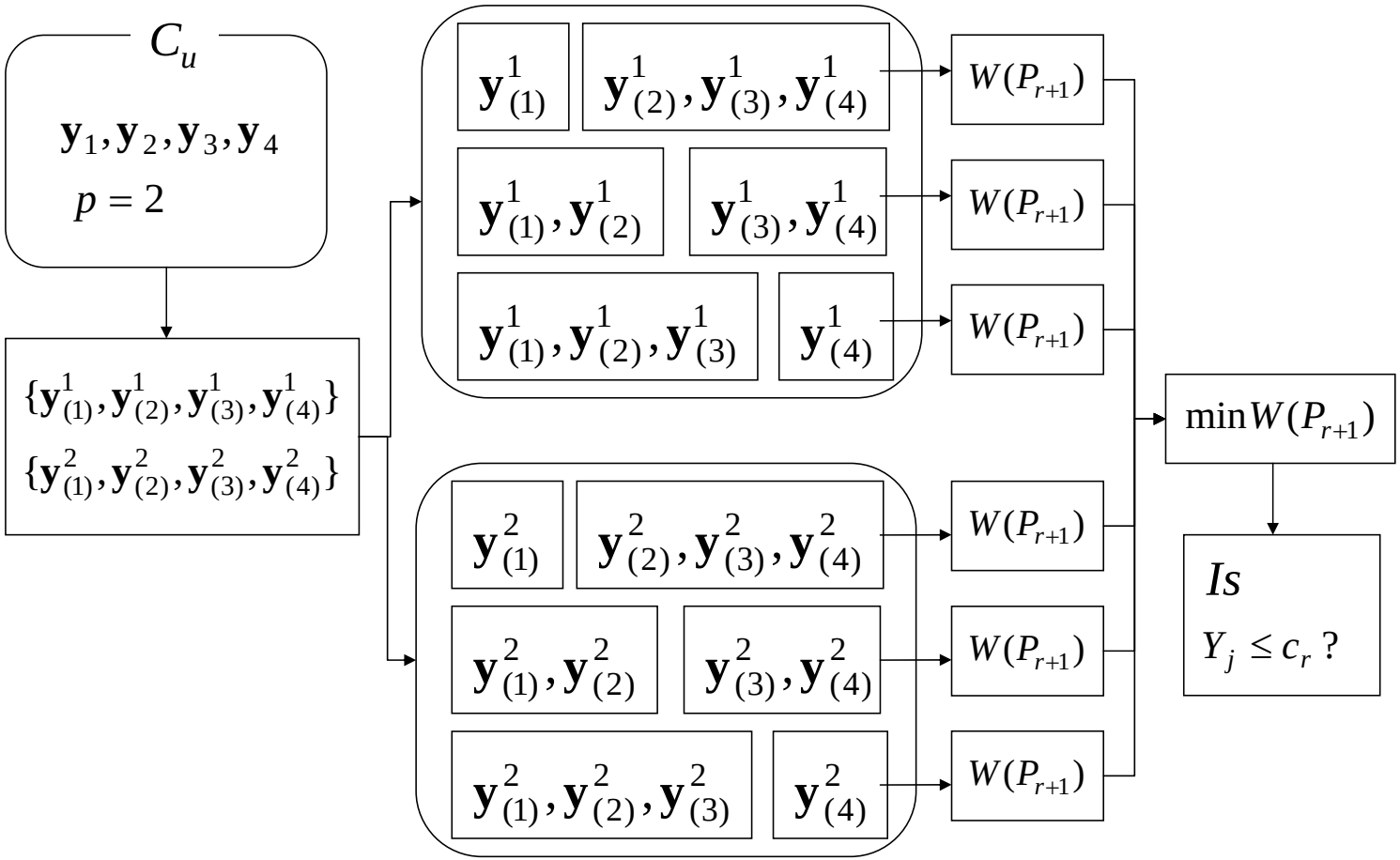


Monothetic Algorithm for Symbolic Objects (Cont.)



Cut point for histogram data: $c_r = M^*_{(q \cup q+1)j}$, or $c_r = S^*_{(q \cup q+1)j}$.

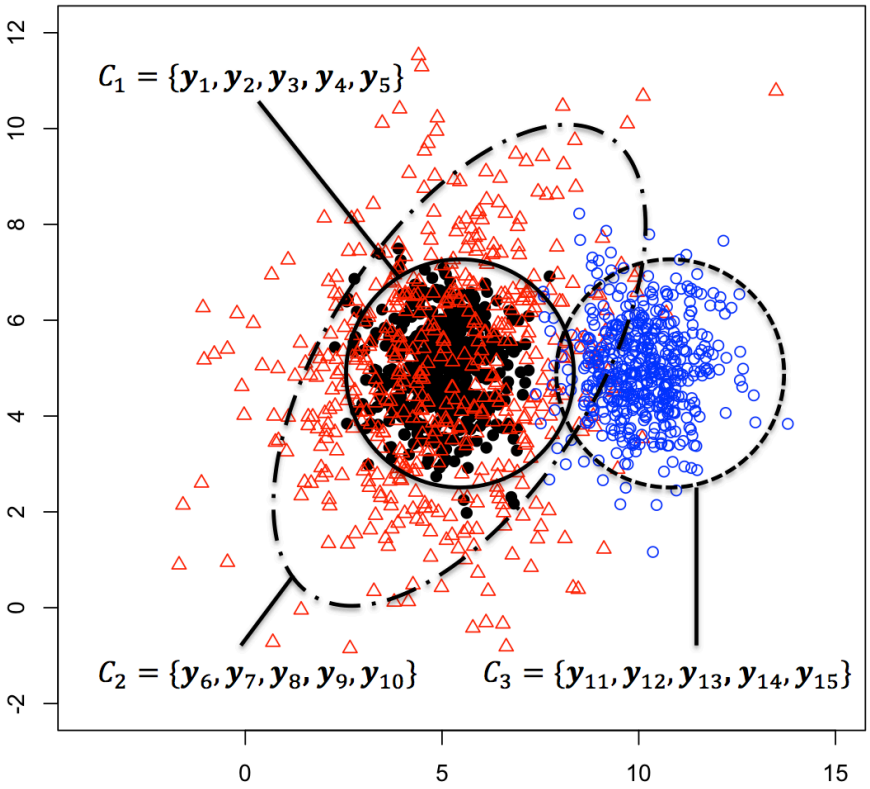
Monothetic Algorithm for Symbolic Objects (Cont.)



Cut point for histogram data: $c_r = M^*_{(q \cup q+1)j}$, or $c_r = S^*_{(q \cup q+1)j}$.

Extends Chavent **(1997, 1998, 2000)**

Simulation

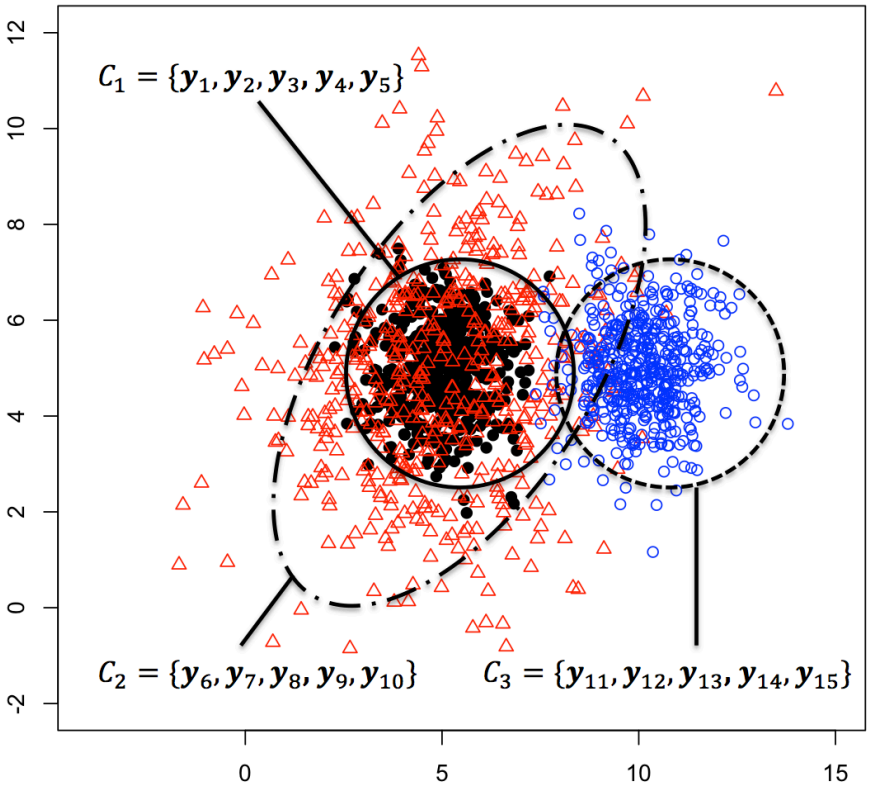


$$BVN \sim \left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

$$BVN \sim \left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 & .8 \\ .8 & 5 \end{bmatrix} \right)$$

$$BVN \sim \left(\begin{bmatrix} 10 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

Simulation



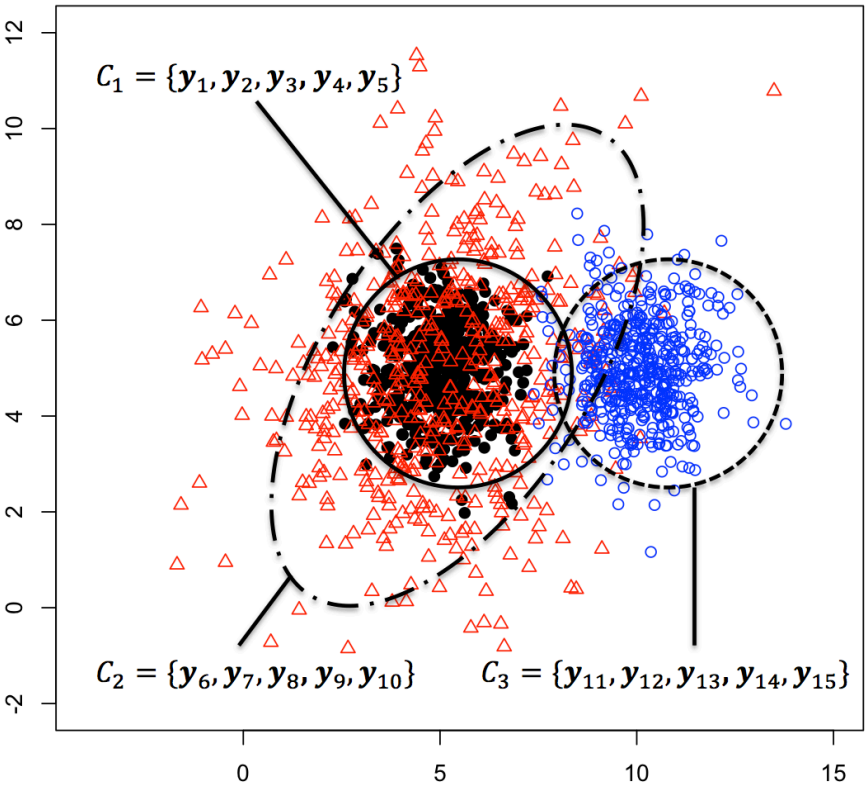
$$BVN \sim \left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

$$BVN \sim \left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 & .8 \\ .8 & 5 \end{bmatrix} \right)$$

$$BVN \sim \left(\begin{bmatrix} 10 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

For 1000 simulated sets,
Brito and Chavent **(2012):** 29 correct;
.... present algorithm: 1000 correct

Simulation



$$BVN \sim \left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

$$BVN \sim \left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 & .8 \\ .8 & 5 \end{bmatrix} \right)$$

$$BVN \sim \left(\begin{bmatrix} 10 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

For 1000 simulated sets,

Brito and Chavent **(2012):** 29 correct;

.... present algorithm: 1000 correct ✓

US Households

American Community Survey Public Use Microdata Sample
House property values (by # bedrooms):

$Y_1 = \leq 2$ bedrooms

$Y_2 = 3$ bedrooms

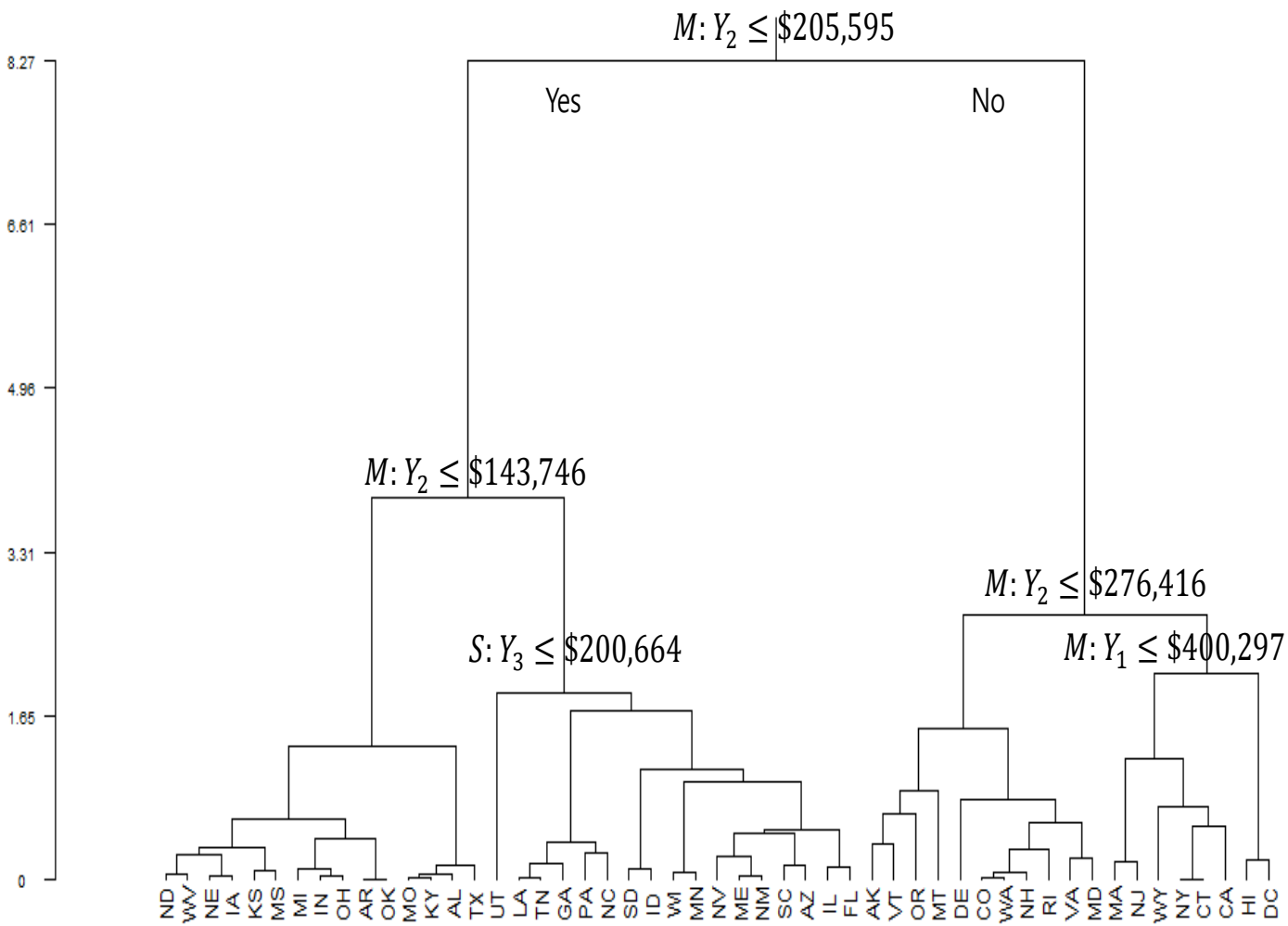
$Y_3 = 4$ bedrooms

$Y_1 = \geq 5$ bedrooms

In all, there are

- 753917 households
- aggregated by state
- gives 51 histograms for each of $p = 4$ variables $(Y_1, \ldots, Y_4)$

US Households



Euclidean Distances

Hvala $\sim\sim$ Thankyou

Regression-based Clustering - Interval Data

Faces data set

Face	Y	X	Face	Y	X
1	[155.00, 157.00]	[58.00, 61.01]	15	[169.03, 170.11]	[59.01, 65.01]
2	[154.00, 160.01]	[57.00, 64.00]	16	[149.34, 155.54]	[54.15, 59.14]
3	[154.01, 161.00]	[57.00, 63.00]	17	[149.34, 155.32]	[52.04, 58.22]
4	[168.86, 172.84]	[58.55, 63.39]	18	[150.33, 157.26]	[52.09, 60.21]
5	[169.85, 175.03]	[60.21, 64.38]	19	[152.64, 157.62]	[51.35, 56.22]
6	[168.76, 175.15]	[61.40, 63.51]	20	[154.64, 157.62]	[52.24, 56.32]
7	[155.26, 160.45]	[53.15, 60.21]	21	[154.83, 157.81]	[50.36, 55.23]
8	[156.26, 161.31]	[51.09, 60.07]	22	[163.08, 167.07]	[66.03, 68.07]
9	[154.47, 160.31]	[55.08, 59.03]	23	[164.00, 168.03]	[65.03, 68.12]
10	[164.00, 168.00]	[55.01, 60.03]	24	[161.01, 167.00]	[64.07, 69.01]
11	[163.00, 170.00]	[54.04, 59.00]	25	[167.15, 171.24]	[64.07, 68.07]
12	[164.01, 169.01]	[55.00, 59.01]	26	[168.15, 172.14]	[63.13, 68.07]
13	[167.11, 171.19]	[61.03, 65.01]	27	[167.11, 171.19]	[63.13, 68.03]
14	[169.14, 173.18]	[60.07, 65.07]			

Regression-based Clustering - Interval Data

Interval data – regression-based partitioning

Classical data: regression approach

Charles **(1977)**, Späth **(1979, 1981, 1982)**, Tibshrani, Walther and Hastie **(2001)**, Shao and Wu **(2005)**, Rao, Wu and Shao **(2007)**, Qian and Wu **(2011)**, DeSarbo and Cron **(1988)**, Zhang **(2003)**, Bougeard, Cariou, Saporta and Niang **(2017)**, Bougeard, Adbi, Saporta and Niang **(2018)**, ...

Regression – interval data

Billard and Diday **(2000,2002)**, deCarvalho, Saporta and Queiroz **(2010)**, Neto and deCarvalho **(2008, 2010)**, Neto, deCarvalho and Freire **(2005)**, Sun and Li **(2014)**, Xu **(2014)**, ...

k-means, adaptive *k*-means, algorithm:

MacQueen **(1967)**, Diday **(1971)**, Diday and Simon **(1976)**, Chavent and Lechevallier **(2002)**, deCarvalho **(2004)**, deSouza and deCarvalho **(2004)**, deSouza et al. **(2004)**, deCarvalho, Brito and Bock **(2006)**, ...

Batagelj, Kejžar and Korenjak-Černe **(2015)**, Irpino, Verde and Lechevallier **(2006)**, Verde and Irpino **(2007)**, Korenjak-Černe, Batagelj and Pavešić **(2011)**, Košmelj and Billard **(2012)**, ...