

Statistical Learning Methods for Process Data

Jingchen Liu

Department of Statistics
Columbia University

July 19, 2021

Paper-pencil tests

Trigonometry

1. Solve for θ . $0 \leq \theta < 2\pi$

a) $2 \cos^2 \theta - 1 = 0$

$$\cos^2 \theta = \frac{1}{2} \Rightarrow \cos \theta = \pm \frac{1}{\sqrt{2}}$$

$$\text{on } [0, 2\pi): \theta = \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}$$

b) $3 \tan^2 \theta - 1 = 0$

$$\tan^2 \theta = \frac{1}{3} \Rightarrow \tan \theta = \pm \frac{1}{\sqrt{3}}$$

$$\text{on } [0, 2\pi): \theta = \frac{\pi}{6}, \frac{5\pi}{6}, \frac{7\pi}{6}, \frac{11\pi}{6}$$

Process Data

Education & Skills Online

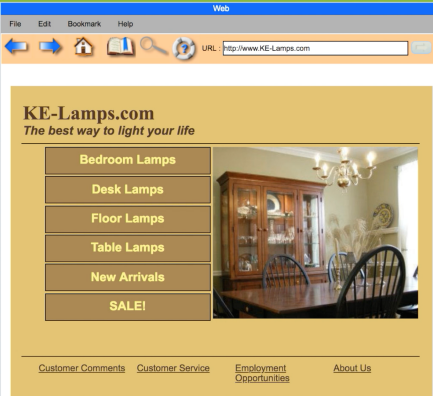
Unit 6

You ordered a desk lamp from KE-Lamps.com.

The desk lamp arrived, but it was not the color you ordered.

Using the company's website, arrange to exchange the lamp you received for the one you ordered.

Once you have finished, click Next to go on.



Event	Time
Start	0.0
Click_CS	2.9
Click_ObtNo	12.1
Button_ObtNo	16.3
Auth_No_Close	18.2
Email	21.4
Web	26.8
Back	28.1
Click_RF	33.3
Reason_Wrong	36.3
Combobox1	39.4
On_AuthBox	40.8
ASCII_7	41.7
ASCII_8	42.2
ASCII_3	43.1
ASCII_4	43.8
Off_AuthBox	44.9
Submit	50.6
Submit_Close	53.5
Next	60.4
Next_OK	62.2

Paper-pencil test, standard test, computer-based interactive test

Trigonometry

1. Solve for θ , $0 \leq \theta < 2\pi$

a) $2 \cos^2 \theta - 1 = 0$

$$\cos^2 \theta = \frac{1}{2} \Rightarrow \cos \theta = \pm \frac{1}{\sqrt{2}}$$

$$\theta = \left\{ 0, 2\pi \right\} \cup \left\{ \frac{\pi}{4}, \frac{7\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4} \right\}$$

b) $3 \tan^2 \theta = 150$

$$\tan^2 \theta = \frac{50}{3} \Rightarrow \tan \theta = \pm \sqrt{\frac{50}{3}}$$

$$\theta = \left\{ 0, 2\pi \right\} \cup \left\{ \frac{\pi}{2}, \frac{3\pi}{2}, \frac{\pi}{2} + \frac{\pi}{\sqrt{150}}, \frac{3\pi}{2} + \frac{\pi}{\sqrt{150}} \right\}$$

高等學校招生
全國統一考試科目答題卡

考號	姓名	考場	考位	科目	分數	備註
01	張三	01	01	數學	85	
02	李四	01	02	數學	78	
03	王五	01	03	數學	92	
04	趙六	01	04	數學	65	
05	孫七	01	05	數學	70	
06	周八	01	06	數學	88	
07	吳九	01	07	數學	75	
08	鄭十	01	08	數學	80	
09	陳十一	01	09	數學	72	
10	林十二	01	10	數學	82	
11	黃十三	01	11	數學	77	
12	周十四	01	12	數學	85	
13	吳十五	01	13	數學	70	
14	鄭十六	01	14	數學	75	
15	陳十七	01	15	數學	80	
16	林十八	01	16	數學	72	
17	黃十九	01	17	數學	82	
18	周二十	01	18	數學	77	
19	吳二十一	01	19	數學	85	
20	鄭二十二	01	20	數學	70	
21	陳二十三	01	21	數學	75	
22	林二十四	01	22	數學	80	
23	黃二十五	01	23	數學	72	
24	周二十六	01	24	數學	82	
25	吳二十七	01	25	數學	77	
26	鄭二十八	01	26	數學	85	
27	陳二十九	01	27	數學	70	
28	林三十	01	28	數學	75	
29	黃三十一	01	29	數學	80	
30	周三十二	01	30	數學	72	
31	吳三十三	01	31	數學	82	
32	鄭三十四	01	32	數學	77	
33	陳三十五	01	33	數學	85	
34	林三十六	01	34	數學	70	
35	黃三十七	01	35	數學	75	
36	周三十八	01	36	數學	80	
37	吳三十九	01	37	數學	72	
38	鄭四十	01	38	數學	82	
39	陳四十一	01	39	數學	77	
40	林四十二	01	40	數學	85	
41	黃四十三	01	41	數學	70	
42	周四十四	01	42	數學	75	
43	吳四十五	01	43	數學	80	
44	鄭四十六	01	44	數學	72	
45	陳四十七	01	45	數學	82	
46	林四十八	01	46	數學	77	
47	黃四十九	01	47	數學	85	
48	周五十	01	48	數學	70	
49	吳五十一	01	49	數學	75	
50	鄭五十二	01	50	數學	80	
51	陳五十三	01	51	數學	72	
52	林五十四	01	52	數學	82	
53	黃五十五	01	53	數學	77	
54	周五十六	01	54	數學	85	
55	吳五十七	01	55	數學	70	
56	鄭五十八	01	56	數學	75	
57	陳五十九	01	57	數學	80	
58	林六十	01	58	數學	72	
59	黃六十一	01	59	數學	82	
60	周六十二	01	60	數學	77	
61	吳六十三	01	61	數學	85	
62	鄭六十四	01	62	數學	70	
63	陳六十五	01	63	數學	75	
64	林六十六	01	64	數學	80	
65	黃六十七	01	65	數學	72	
66	周六十八	01	66	數學	82	
67	吳六十九	01	67	數學	77	
68	鄭七十	01	68	數學	85	
69	陳七十一	01	69	數學	70	
70	林七十二	01	70	數學	75	
71	黃七十三	01	71	數學	80	
72	周七十四	01	72	數學	72	
73	吳七十五	01	73	數學	82	
74	鄭七十六	01	74	數學	77	
75	陳七十七	01	75	數學	85	
76	林七十八	01	76	數學	70	
77	黃七十九	01	77	數學	75	
78	周八十	01	78	數學	80	
79	吳八十一	01	79	數學	72	
80	鄭八十二	01	80	數學	82	
81	陳八十三	01	81	數學	77	
82	林八十四	01	82	數學	85	
83	黃八十五	01	83	數學	70	
84	周八十六	01	84	數學	75	
85	吳八十七	01	85	數學	80	
86	鄭八十八	01	86	數學	72	
87	陳八十九	01	87	數學	82	
88	林九十	01	88	數學	77	
89	黃九十一	01	89	數學	85	
90	周九十二	01	90	數學	70	
91	吳九十三	01	91	數學	75	
92	鄭九十四	01	92	數學	80	
93	陳九十五	01	93	數學	72	
94	林九十六	01	94	數學	82	
95	黃九十七	01	95	數學	77	
96	周九十八	01	96	數學	85	
97	吳九十九	01	97	數學	70	
98	鄭一百	01	98	數學	75	
99	陳一百零一	01	99	數學	80	
100	林一百零二	01	100	數學	72	

Education & Skills Online

UPSE

You ordered a Desk Lamp from KE-Lamps.com

The desk lamp arrived, but it was not the color you ordered.

Using the company's website, arrange to exchange the lamp you received for the one you ordered.

Once you have finished, click Next to go on.

KE-Lamps.com
The best way to light your life

- Bedroom Lamps
- Desk Lamps
- Floor Lamps
- Table Lamps
- New Arrivals
- SALE!

Customer Comments Customer Service Employee's Accounts About Us

Process response



Process response



Process data



Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ **Existing problems:**
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ Existing problems:
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ Existing problems:
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ Existing problems:
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ **Existing problems:**
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ **Existing problems:**
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ **Existing problems:**
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ **Existing problems:**
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ **Existing problems:**
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Process Data Research

- ▶ **New problems:**
 - ▶ problem-solving strategy analysis
 - ▶ cognitive structures
 - ▶ ...

- ▶ **Existing problems:**
 - ▶ assessment
 - ▶ differential item functioning
 - ▶ computerized adaptive testing
 - ▶ adaptive learning, etc.
 - ▶ ...

Four aspects

- ▶ **Research objective**
- ▶ Empirical findings
- ▶ Technical details: references
<http://www.scientifichpc.com/processdata/pub.html>
- ▶ Implementation: sessions 3 and 4

Four aspects

- ▶ **Research objective**
- ▶ **Empirical findings**
- ▶ Technical details: references
<http://www.scientifichpc.com/processdata/pub.html>
- ▶ **Implementation: sessions 3 and 4**

Four aspects

- ▶ Research objective
- ▶ Empirical findings
- ▶ Technical details: references
<http://www.scientifchpc.com/processdata/pub.html>
- ▶ Implementation: sessions 3 and 4

Four aspects

- ▶ **Research objective**
- ▶ **Empirical findings**
- ▶ Technical details: references
<http://www.scientifchpc.com/processdata/pub.html>
- ▶ **Implementation**: sessions 3 and 4

Content of the overview



- ▶ Feature extraction: $(\theta_1, \dots, \theta_K) \in \mathbb{R}^K$
- ▶ Partial scoring
- ▶ Removing differential item functioning

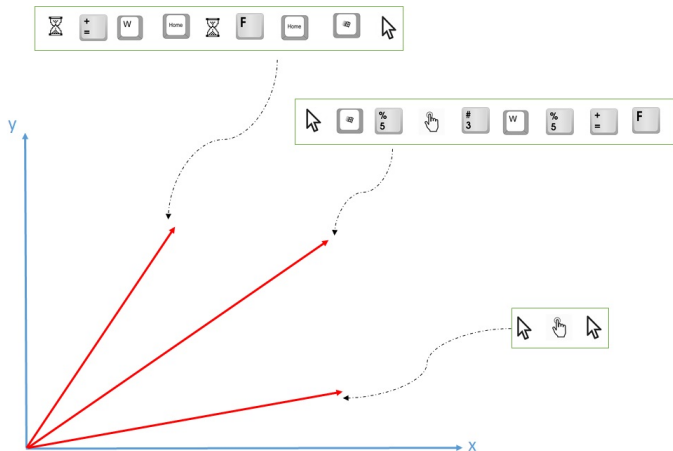
Content of the overview



- ▶ Feature extraction: $(\theta_1, \dots, \theta_K) \in \mathbb{R}^K$
- ▶ Partial scoring
- ▶ Removing differential item functioning

Feature Extraction

Embedding



Embedding



- ▶ **Process** response:

```
action:Start, Click_cs, Click_ObtNo, ..., Next, Next_OK
time: 0.0 , 2.9 , 12.1 , ..., 60.4, 62.2
```

- ▶ Summarize the **process** to $(\theta_1, \dots, \theta_k) \in \mathbb{R}^k$.

A large variety of items

- ▶ Email handling/classification, spread sheet handling, scheduling, web browsing/comprehension, etc.

- ▶ Learning/interactive with a system, designing experiments, etc.

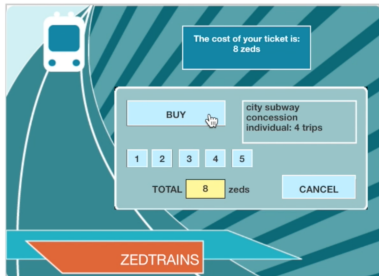
Process Data

TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- o Choose the train network you want (subway or country).
- o Choose the type of fare (full or concession).
- o Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.



Question TICKETS

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY.

Once you have pressed BUY, you cannot return to the question.

Very noisy

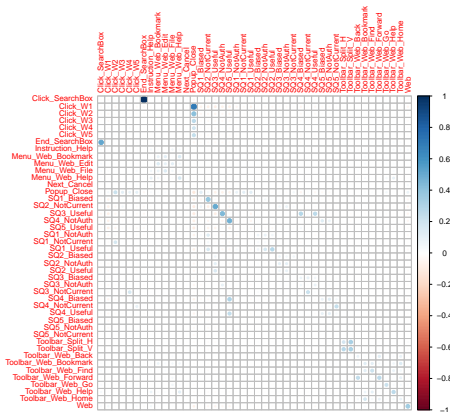


Figure: Lag 1 autocorrelation: $\text{cor}(a_t, a_{t+1})$

Very noisy

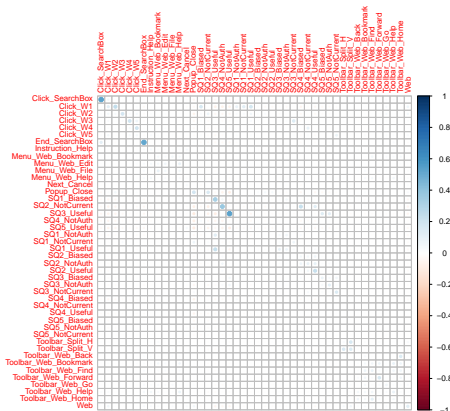


Figure: Lag 2 autocorrelation: $\text{cor}(a_t, a_{t+2})$

Objective

- ▶ **Denoising**: aggregate the process to strengthen the signal.
- ▶ **Formatting**: $\theta \in \mathbb{R}^K$, for $K \geq 100$.
- ▶ **Dimension reduction**.

Latent structure extraction



a_1 a_2 a_3

$a_j \in \mathcal{A} = \{1, \dots, m\}$

- ▶ Process length varies among individuals in the range of [3,1000]
- ▶ The number of possible actions m varies among items in the range of [20, 300].
- ▶ $(a_1, \dots, a_{n_i}) \Rightarrow (\theta_1, \dots, \theta_k)$

Evaluation criteria

- ▶ Process features: $(a_1, \dots, a_{n_i}) \Rightarrow \theta = (\theta_1, \dots, \theta_k)$
- ▶ Benchmark: $(a_1, \dots, a_{n_i}) \Rightarrow r \in \{\checkmark, \times\}$: task accomplishment
- ▶ Does θ contain more information than r ? How much information do we lose?

Assessing latent variable

- ▶ y : a different variable, such as literacy score.
 - ▶ \hat{y}_θ : prediction based on θ
 - versus
 - ▶ \hat{y}_r : prediction based on r

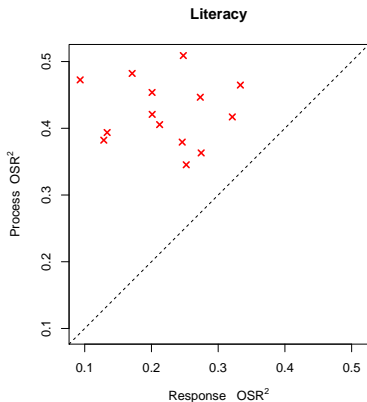


Figure: $cor^2(y, \hat{y}_r)$ versus $cor^2(y, \hat{y}_\theta)$

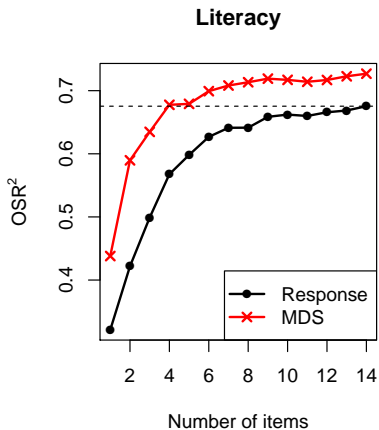


Figure: $cor^2(y, \hat{y}_{r_1, \dots, r_k})$ versus $cor^2(y, \hat{y}_{\theta_1, \dots, \theta_k})$

Latent structure extraction

- ▶ Multidimensional scaling

Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2019) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*.

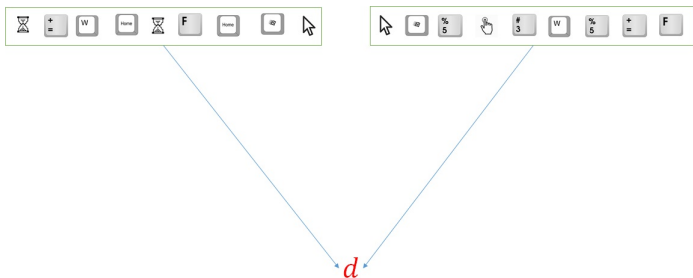
- ▶ Autoencoder

Tang, X., Wang, Z., Liu, J., and Ying, Z. (2019) An Exploratory Analysis of the Latent Structure of Process Data via Action Sequence Autoencoder. *British Journal of Mathematical and Statistical Psychology*.

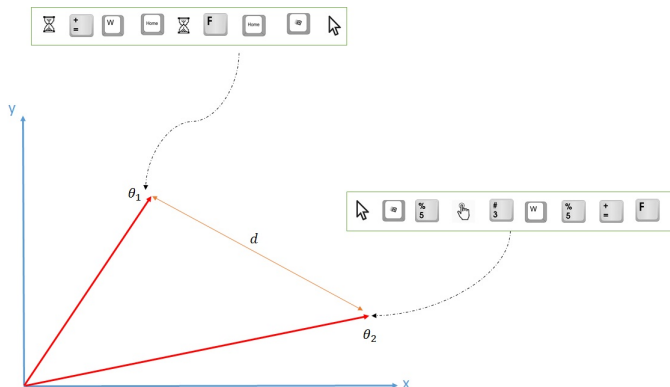
- ▶ R package

Tang, X., Zhang, S., Wang, Z., Liu, J., and Ying, Z. (2021) ProcData: An R Package for Process Data Analysis. *Psychometrika*. To appear.

Multidimensional scaling



Multidimensional scaling



Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_i})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_j})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij} = d(\mathbf{a}_i, \mathbf{a}_j)$$

- ▶ The distance Gómez-Alonso and Valls (2008)

$$d(\mathbf{a}_i, \mathbf{a}_j) = \frac{f(\mathbf{a}_i, \mathbf{a}_j) + g(\mathbf{a}_i, \mathbf{a}_j)}{n_i + n_j},$$

- ▶ Common actions:

$$f(\mathbf{a}_i, \mathbf{a}_j) = \frac{\sum_{a \in C_{ij}} \sum_{k=1}^{K_{ij}^a} |s_i^a(k) - s_j^a(k)|}{\max\{n_i, n_j\}},$$

- ▶ Uncommon actions:

$$g(\mathbf{a}_i, \mathbf{a}_j) = \sum_{a \in U_{ij}} n_i^a + \sum_{a \in U_{ji}} n_j^a,$$

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_i})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_j})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij} = d(\mathbf{a}_i, \mathbf{a}_j)$$

- ▶ The distance Gómez-Alonso and Valls (2008)

$$d(\mathbf{a}_i, \mathbf{a}_j) = \frac{f(\mathbf{a}_i, \mathbf{a}_j) + g(\mathbf{a}_i, \mathbf{a}_j)}{n_i + n_j},$$

- ▶ Common actions:

$$f(\mathbf{a}_i, \mathbf{a}_j) = \frac{\sum_{a \in C_{ij}} \sum_{k=1}^{K_{ij}^a} |s_i^a(k) - s_j^a(k)|}{\max\{n_i, n_j\}},$$

- ▶ Uncommon actions:

$$g(\mathbf{a}_i, \mathbf{a}_j) = \sum_{a \in U_{ij}} n_i^a + \sum_{a \in U_{ji}} n_j^a,$$

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_i})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_j})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij} = d(\mathbf{a}_i, \mathbf{a}_j)$$

- ▶ The distance Gómez-Alonso and Valls (2008)

$$d(\mathbf{a}_i, \mathbf{a}_j) = \frac{f(\mathbf{a}_i, \mathbf{a}_j) + g(\mathbf{a}_i, \mathbf{a}_j)}{n_i + n_j},$$

- ▶ Common actions:

$$f(\mathbf{a}_i, \mathbf{a}_j) = \frac{\sum_{a \in C_{ij}} \sum_{k=1}^{K_{ij}^a} |s_i^a(k) - s_j^a(k)|}{\max\{n_i, n_j\}},$$

- ▶ Uncommon actions:

$$g(\mathbf{a}_i, \mathbf{a}_j) = \sum_{a \in U_{ij}} n_i^a + \sum_{a \in U_{ji}} n_j^a,$$

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_i})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_j})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij} = d(\mathbf{a}_i, \mathbf{a}_j)$$

- ▶ The distance Gómez-Alonso and Valls (2008)

$$d(\mathbf{a}_i, \mathbf{a}_j) = \frac{f(\mathbf{a}_i, \mathbf{a}_j) + g(\mathbf{a}_i, \mathbf{a}_j)}{n_i + n_j},$$

- ▶ Common actions:

$$f(\mathbf{a}_i, \mathbf{a}_j) = \frac{\sum_{a \in C_{ij}} \sum_{k=1}^{K_{ij}^a} |s_i^a(k) - s_j^a(k)|}{\max\{n_i, n_j\}},$$

- ▶ Uncommon actions:

$$g(\mathbf{a}_i, \mathbf{a}_j) = \sum_{a \in U_{ij}} n_i^a + \sum_{a \in U_{ji}} n_j^a,$$

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_1})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_2})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij}, \quad 1 \leq i, j \leq n$$

- ▶ Distance matrices $D = (d_{ij})_{n \times n}$

- ▶ Latent variable $\theta_i \in \mathbb{R}^k$.

$$d_{ij} = \varphi(\theta_i, \theta_j) + \varepsilon_{ij}$$

where $\varphi(\theta_i, \theta_j) = |\theta_i - \theta_j|$.

- ▶ Optimization

$$\min_{\theta_1, \dots, \theta_n} \sum_{i < j} |d_{ij} - \varphi(\theta_i, \theta_j)|^2.$$

- ▶ Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2019) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*.

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_1})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_2})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij}, \quad 1 \leq i, j \leq n$$

- ▶ Distance matrices $D = (d_{ij})_{n \times n}$

- ▶ Latent variable $\theta_i \in \mathbb{R}^k$.

$$d_{ij} = \varphi(\theta_i, \theta_j) + \varepsilon_{ij}$$

where $\varphi(\theta_i, \theta_j) = |\theta_i - \theta_j|$.

- ▶ Optimization

$$\min_{\theta_1, \dots, \theta_n} \sum_{i < j} |d_{ij} - \varphi(\theta_i, \theta_j)|^2.$$

- ▶ Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2019) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*.

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_1})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_2})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij}, \quad 1 \leq i, j \leq n$$

- ▶ Distance matrices $D = (d_{ij})_{n \times n}$

- ▶ Latent variable $\theta_i \in \mathbb{R}^k$.

$$d_{ij} = \varphi(\theta_i, \theta_j) + \varepsilon_{ij}$$

where $\varphi(\theta_i, \theta_j) = |\theta_i - \theta_j|$.

- ▶ Optimization

$$\min_{\theta_1, \dots, \theta_n} \sum_{i < j} |d_{ij} - \varphi(\theta_i, \theta_j)|^2.$$

- ▶ Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2019) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*.

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_1})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_2})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij}, \quad 1 \leq i, j \leq n$$

- ▶ Distance matrices $D = (d_{ij})_{n \times n}$

- ▶ Latent variable $\theta_i \in \mathbb{R}^k$.

$$d_{ij} = \varphi(\theta_i, \theta_j) + \varepsilon_{ij}$$

where $\varphi(\theta_i, \theta_j) = |\theta_i - \theta_j|$.

- ▶ Optimization

$$\min_{\theta_1, \dots, \theta_n} \sum_{i < j} |d_{ij} - \varphi(\theta_i, \theta_j)|^2.$$

- ▶ Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2019) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*.

Multidimensional scaling

- ▶ Two response processes: $\mathbf{a}_i = (a_{i1}, \dots, a_{in_1})$, $\mathbf{a}_j = (a_{j1}, \dots, a_{jn_2})$

$$(\mathbf{a}_i, \mathbf{a}_j) \rightarrow d_{ij}, \quad 1 \leq i, j \leq n$$

- ▶ Distance matrices $D = (d_{ij})_{n \times n}$

- ▶ Latent variable $\theta_i \in \mathbb{R}^k$.

$$d_{ij} = \varphi(\theta_i, \theta_j) + \varepsilon_{ij}$$

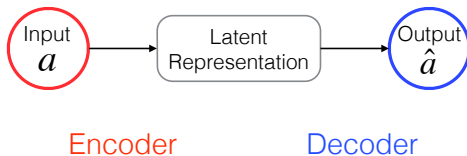
where $\varphi(\theta_i, \theta_j) = |\theta_i - \theta_j|$.

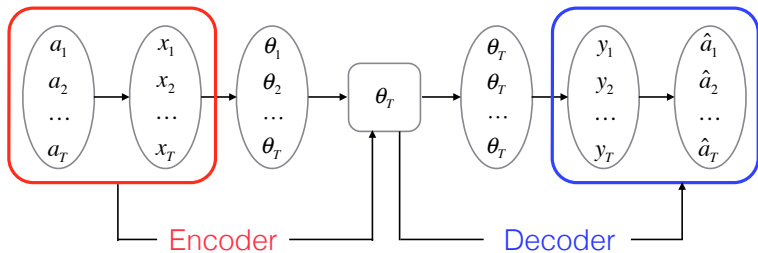
- ▶ Optimization

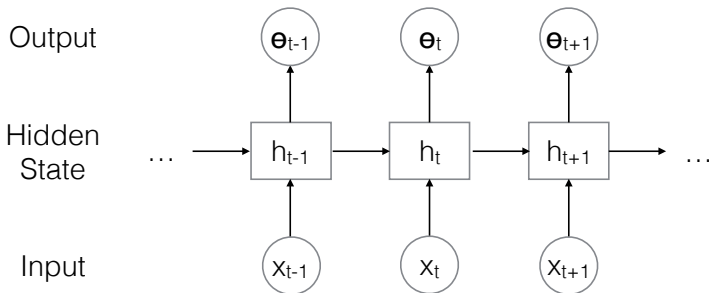
$$\min_{\theta_1, \dots, \theta_n} \sum_{i < j} |d_{ij} - \varphi(\theta_i, \theta_j)|^2.$$

- ▶ Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2019) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*.

Autoencoder







Autoencoder

- ▶ Autoencoder via tensorflow
- ▶ Tang, X., Wang, Z., Liu, J., and Ying, Z. (2019) An Exploratory Analysis of the Latent Structure of Process Data via Action Sequence Autoencoder. *British Journal of Mathematical and Statistical Psychology*.

Criterion

- ▶ $\mathbf{a} \Rightarrow \theta \in \mathbb{R}^K$: multidimensional scaling or autoencoder
- ▶ $\mathbf{a} \Rightarrow r \in \{\checkmark, \times\}$: task accomplishment

Assessing latent variable

- ▶ How much information did we lose to r ?
- ▶ \hat{r}_θ : prediction of task accomplishment based on θ .
- ▶ To what extent θ captures the information in r .

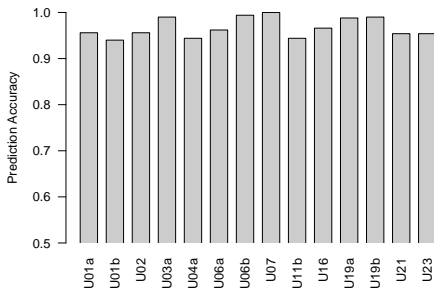


Figure: $P(r = \hat{r}_\theta)$ based on MDS

Assessing latent variable through prediction

- ▶ How much additional information do we gain?
- ▶ y : a different variable, such as numeracy score.
 - ▶ \hat{y}_θ : prediction based on θ

Versus

- ▶ \hat{y}_r : prediction based on r

$$\text{cor}(y, \hat{y})$$

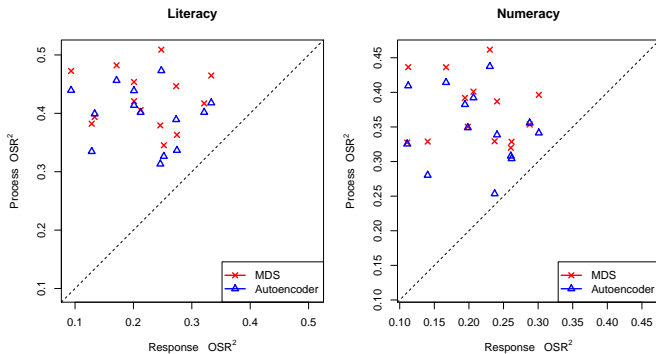


Figure: \hat{y}_θ versus \hat{y}_r

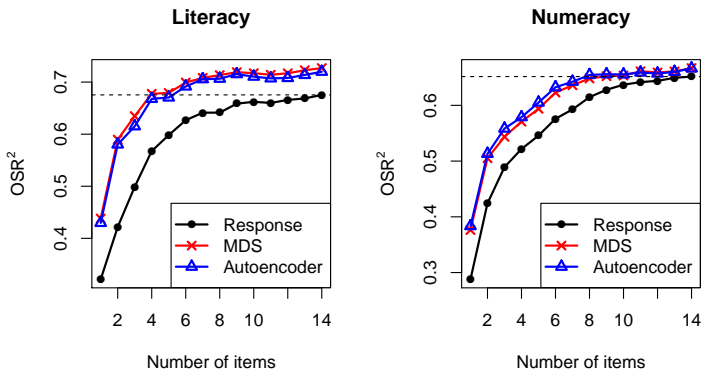


Figure: $\hat{y}(\theta_1, \dots, \theta_k)$ versus $\hat{y}(r_1, \dots, r_k)$

Demographic variables

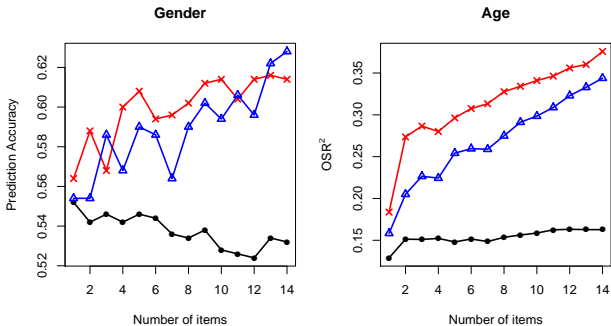


Figure: Prediction of age and gender

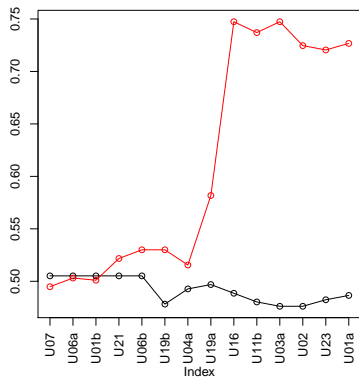


Figure: Prediction of country

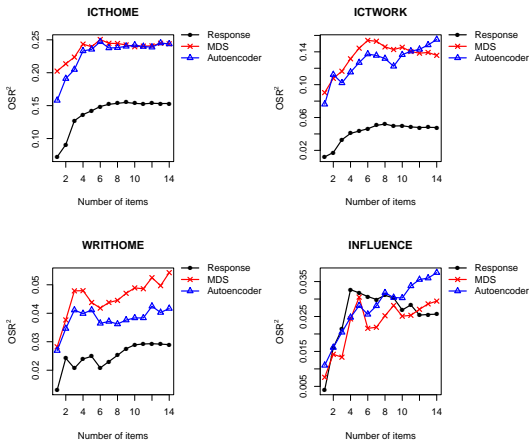


Figure: ICTHOME,ICTWORK,WRITHOME,INFLUENCE

Closer look at the latent variables – principle components

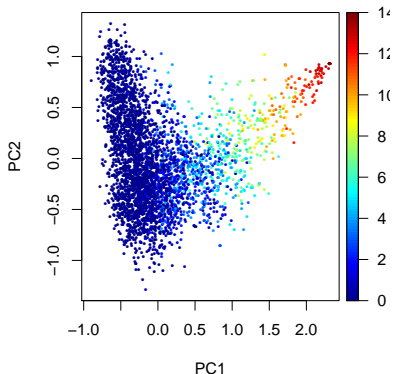


Figure: Principle components of θ . PC1 correlation with number of skipped items: 0.88

Latent variable in subpopulations

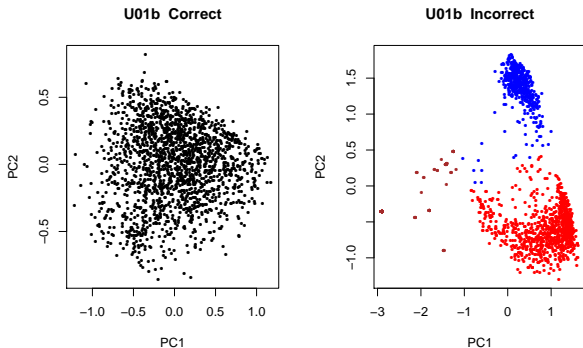
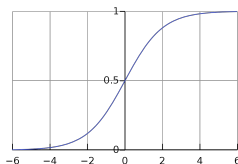


Figure: Correct vs incorrect

Applications

- ▶ Partial score: improving assessment accuracy
- ▶ Removing differential item functioning (DIF)

Partial scores



- ▶ IRT model

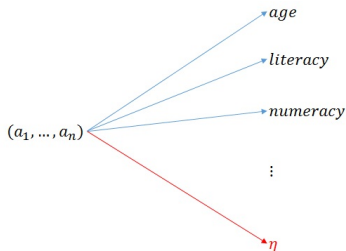
$$P(r_j = 1|\eta) = \frac{e^{a_j(\eta - b_j)}}{1 + e^{a_j(\eta - b_j)}}$$

- ▶ Assessment (grading/scoring) via maximum likelihood estimate

$$\hat{\eta}(r_1, \dots, r_J) = \arg \max_{\eta} \prod_j P(r_j|\eta)$$

- ▶ Process-data-based scores: $\hat{\eta}(a_j)$

Partial scores



- ▶ Difficulty: **validity**
- ▶ **Partial scores** based on the entire response process
- ▶ Scores guided by y .
- ▶ Generalization: assisting any measurable characteristics

Partial scores guided by response

- ▶ Train a scoring rule $f(\mathbf{a})$ towards η .
- ▶ $r = g(\eta) + \varepsilon \sim \mathbf{a}$, ε is \mathbf{a} -predictable.
- ▶ $g(\eta) + \varepsilon \sim \mathbf{a}$
- ▶ Proxies
 - ▶ $\hat{\eta}(r_{-j}) = \eta + \varepsilon$
 - ▶ \mathbf{a} : process features θ

$$\hat{\eta}(r_{-j}) \sim \theta$$

Partial scores guided by response

- ▶ Train a scoring rule $f(\mathbf{a})$ towards η .
- ▶ $r = g(\eta) + \varepsilon \sim \mathbf{a}$, ε is \mathbf{a} -predictable.
- ▶ $g(\eta) + \varepsilon \sim \mathbf{a}$
- ▶ Proxies
 - ▶ $\hat{\eta}(r_{-j}) = \eta + \varepsilon$
 - ▶ \mathbf{a} : process features θ

$$\hat{\eta}(r_{-j}) \sim \theta$$

Partial scores guided by response

- ▶ Train a scoring rule $f(\mathbf{a})$ towards η .
- ▶ $r = g(\eta) + \varepsilon \sim \mathbf{a}$, ε is \mathbf{a} -predictable.
- ▶ $g(\eta) + \varepsilon \sim \mathbf{a}$
- ▶ Proxies
 - ▶ $\hat{\eta}(r_{-j}) = \eta + \varepsilon$
 - ▶ \mathbf{a} : process features θ

$$\hat{\eta}(r_{-j}) \sim \theta$$

Partial scores guided by response

- ▶ Train a scoring rule $f(\mathbf{a})$ towards η .
- ▶ $r = g(\eta) + \varepsilon \sim \mathbf{a}$, ε is \mathbf{a} -predictable.
- ▶ $g(\eta) + \varepsilon \sim \mathbf{a}$
- ▶ Proxies
 - ▶ $\hat{\eta}(r_{-j}) = \eta + \varepsilon$
 - ▶ \mathbf{a} : process features θ

$$\hat{\eta}(r_{-j}) \sim \theta$$

Partial scores guided by response

- ▶ Train a scoring rule $f(\mathbf{a})$ towards η .
- ▶ $r = g(\eta) + \varepsilon \sim \mathbf{a}$, ε is \mathbf{a} -predictable.
- ▶ $g(\eta) + \varepsilon \sim \mathbf{a}$
- ▶ Proxies
 - ▶ $\hat{\eta}(r_{-j}) = \eta + \varepsilon$
 - ▶ \mathbf{a} : process features θ

$$\hat{\eta}(r_{-j}) \sim \theta$$

Partial scores guided by response

- ▶ Train a scoring rule $f(\mathbf{a})$ towards η .
- ▶ $r = g(\eta) + \varepsilon \sim \mathbf{a}$, ε is \mathbf{a} -predictable.
- ▶ $g(\eta) + \varepsilon \sim \mathbf{a}$
- ▶ Proxies
 - ▶ $\hat{\eta}(r_{-j}) = \eta + \varepsilon$
 - ▶ \mathbf{a} : process features θ

$$\hat{\eta}(r_{-j}) \sim \theta$$

Partial scores guided by response

1. Extract process feature θ_j for each item
2. Compute IRT score $\eta(r_{-j})$
3. Regression: $\hat{\eta} = f(\theta_j)$
4. Score: $f(\theta_j)$

Evaluation criterion – improving reliability

- ▶ 14 PSTRE items in total
- ▶ Randomization
 - ▶ Session 1 (7 items): $\hat{\eta}_1$ IRT estimate, $\tilde{\eta}_1$ process data estimate
 - ▶ Session 2 (7 items): $\hat{\eta}_2$ IRT estimate
- ▶ Compare $cor(\hat{\eta}_1, \hat{\eta}_2)$ against $cor(\tilde{\eta}_1, \hat{\eta}_2)$

Evaluation criterion – improving reliability

- ▶ 14 PSTRE items in total
- ▶ Randomization
 - ▶ Session 1 (7 items): $\hat{\eta}_1$ IRT estimate, $\tilde{\eta}_1$ process data estimate
 - ▶ Session 2 (7 items): $\hat{\eta}_2$ IRT estimate
- ▶ Compare $cor(\hat{\eta}_1, \hat{\eta}_2)$ against $cor(\tilde{\eta}_1, \hat{\eta}_2)$

Evaluation criterion – improving reliability

- ▶ 14 PSTRE items in total
- ▶ Randomization
 - ▶ Session 1 (7 items): $\hat{\eta}_1$ IRT estimate, $\tilde{\eta}_1$ process data estimate
 - ▶ Session 2 (7 items): $\hat{\eta}_2$ IRT estimate
- ▶ Compare $cor(\hat{\eta}_1, \hat{\eta}_2)$ against $cor(\tilde{\eta}_1, \hat{\eta}_2)$

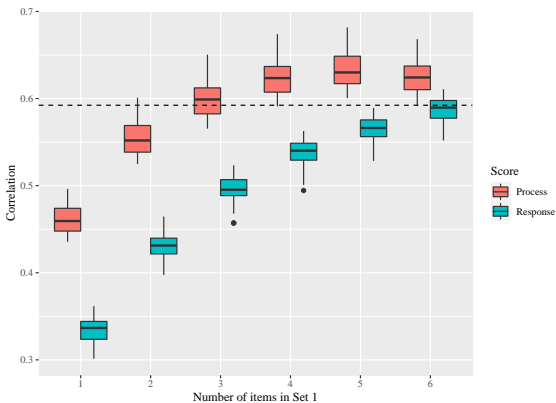
Evaluation criterion – improving reliability

- ▶ 14 PSTRE items in total
- ▶ Randomization
 - ▶ Session 1 (7 items): $\hat{\eta}_1$ IRT estimate, $\tilde{\eta}_1$ process data estimate
 - ▶ Session 2 (7 items): $\hat{\eta}_2$ IRT estimate
- ▶ Compare $cor(\hat{\eta}_1, \hat{\eta}_2)$ against $cor(\tilde{\eta}_1, \hat{\eta}_2)$

Evaluation criterion – improving reliability

- ▶ 14 PSTRE items in total
- ▶ Randomization
 - ▶ Session 1 (7 items): $\hat{\eta}_1$ IRT estimate, $\tilde{\eta}_1$ process data estimate
 - ▶ Session 2 (7 items): $\hat{\eta}_2$ IRT estimate
- ▶ Compare $cor(\hat{\eta}_1, \hat{\eta}_2)$ against $cor(\tilde{\eta}_1, \hat{\eta}_2)$

Out-of-sample Reliability



Differential item functioning (DIF)

- ▶ About differential item functioning
- ▶ Literature: identifying DIF
- ▶ Process data: removing DIF

Framework

- ▶ Response Y with item response function

$$r \sim \eta, x_1, \dots, x_m$$

$$f(r|\eta, x_1, \dots, x_m)$$

- ▶ Assessment model

$$r \sim \eta$$

- ▶ Observed item response function

$$f(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ Two groups: 1 and 2.

$$f_g(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi_g(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ DIF: unbalanced distributions $\pi_g(\eta, x_1, \dots, x_m)$ in groups 1 and 2.

Framework

- ▶ Response Y with item response function

$$r \sim \eta, x_1, \dots, x_m$$

$$f(r|\eta, x_1, \dots, x_m)$$

- ▶ Assessment model

$$r \sim \eta$$

- ▶ Observed item response function

$$f(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ Two groups: 1 and 2.

$$f_g(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi_g(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ DIF: unbalanced distributions $\pi_g(\eta, x_1, \dots, x_m)$ in groups 1 and 2.

Framework

- ▶ Response Y with item response function

$$r \sim \eta, x_1, \dots, x_m$$

$$f(r|\eta, x_1, \dots, x_m)$$

- ▶ Assessment model

$$r \sim \eta$$

- ▶ Observed item response function

$$f(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ Two groups: 1 and 2.

$$f_g(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi_g(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ DIF: unbalanced distributions $\pi_g(\eta, x_1, \dots, x_m)$ in groups 1 and 2.

Framework

- ▶ Response Y with item response function

$$r \sim \eta, x_1, \dots, x_m$$

$$f(r|\eta, x_1, \dots, x_m)$$

- ▶ Assessment model

$$r \sim \eta$$

- ▶ Observed item response function

$$f(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ Two groups: 1 and 2.

$$f_g(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi_g(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ DIF: unbalanced distributions $\pi_g(\eta, x_1, \dots, x_m)$ in groups 1 and 2.

Framework

- ▶ Response Y with item response function

$$r \sim \eta, x_1, \dots, x_m$$

$$f(r|\eta, x_1, \dots, x_m)$$

- ▶ Assessment model

$$r \sim \eta$$

- ▶ Observed item response function

$$f(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ Two groups: 1 and 2.

$$f_g(r|\eta) = \int f(r|\eta, x_1, \dots, x_m)\pi_g(x_1, \dots, x_m|\eta)dx_1 \dots dx_m$$

- ▶ DIF: unbalanced distributions $\pi_g(\eta, x_1, \dots, x_m)$ in groups 1 and 2.

Removing DIF

- ▶ Ideal solution: use the correct item response function $f(r|\eta, x_1, \dots, x_m)$.
- ▶ x_1, \dots, x_m unobserved
- ▶ Use process data features as proxies, $\theta_1, \dots, \theta_K$.

Removing DIF – technical aspects

- ▶ Over fitting: including all process feature

$$f(r|\eta, \theta_1, \dots, \theta_K) = f(r|\theta_1, \dots, \theta_K)$$

- ▶ Variable selection: minimum amount of process data so that

$$\|f_1(r|\eta, \theta_{i_1}, \dots, \theta_{i_j}) - f_2(r|\eta, \theta_{i_1}, \dots, \theta_{i_j})\| \approx 0$$

Why are process features good proxies?

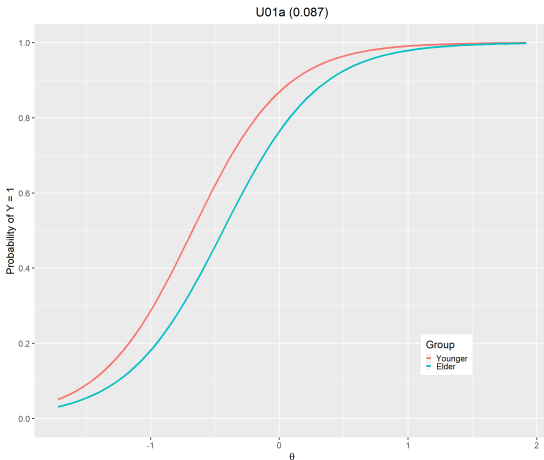
- ▶ Process features contain sufficient information to remove DIF

$$r = f(\theta_1, \dots, \theta_K)$$

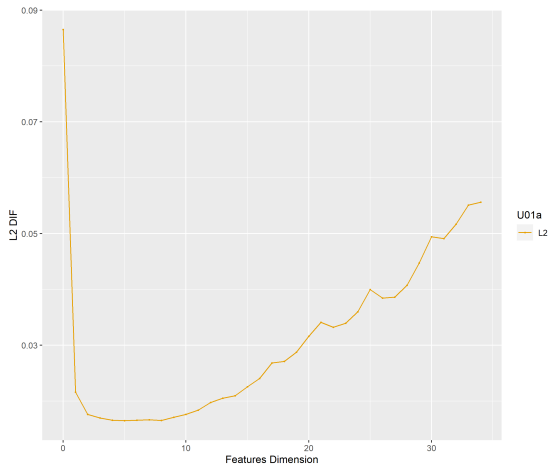
- ▶ Overfitting
- ▶ No extra information to estimate η
- ▶ DIF versus information: include minimum amount of process data to maintain non-differentiation of item functioning.
- ▶ A forward search algorithm

$$\max_j \|f_1(r|\eta, \theta_{i_1}, \dots, \theta_{i_k}, \theta_j) - f_2(r|\eta, \theta_{i_1}, \dots, \theta_{i_k}, \theta_j)\|$$

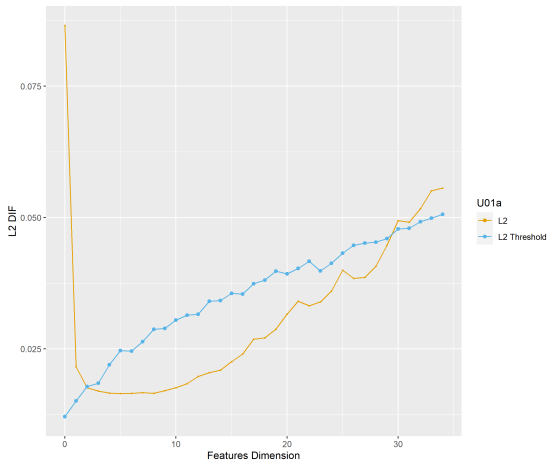
Empirical results



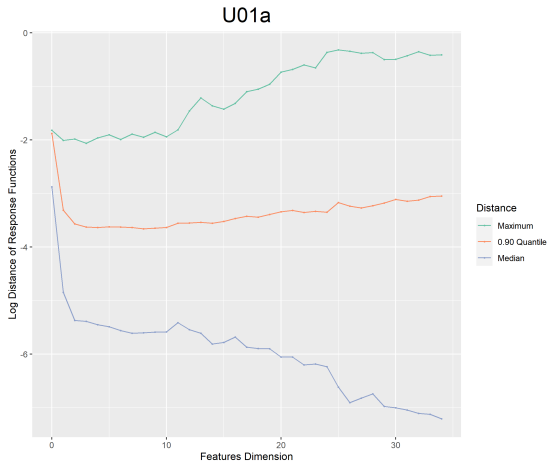
Empirical results



Empirical results



Empirical results



References

- ▶ <http://www.scientifichpc.com/processdata/pub.html>
- ▶ Feature extraction
 - ▶ Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2020) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*.
 - ▶ Tang, X., Wang, Z., Liu, J., and Ying, Z. (2020) An Exploratory Analysis of the Latent Structure of Process Data via Action Sequence Autoencoder. *British Journal of Mathematical and Statistical Psychology*.
 - ▶ Tang, X., Zhang, S., Wang, Z., Liu, J., and Ying, Z. (2021) ProcData: An R Package for Process Data Analysis. *Psychometrika*. To appear.
- ▶ Partial Scoring:
 - ▶ Zhang, S., Wang, Z., Qi, J., Liu, J., and Ying, Z. Accurate Assessment via Process Data.
- ▶ Removing differential item functioning

Acknowledgement

Zhiliang Ying

Zhi Wang, Jitong Qi, Brian Ling

OECD, Educational Testing Service: Qiwei He

National Science Foundation, Army Research Office