順序付き多変量データのための客観的総合指数 ―TextilePlotとの関係―

統計関連学会連合大会（金沢大学）
「データサイエンスの世界的潮流とその展望」
２０１６年９月７日　清 智也

参考文献：Sei (2016) JMVA, 147, 247―164.

目的

- 各変量に「大きい値ほど良い」という順序が与えられた多変量データを考える。

- 例：５科目の成績、十種競技、大学ランキング

- 客観的な総合指数を作りたい。
- 記述統計、教師なし

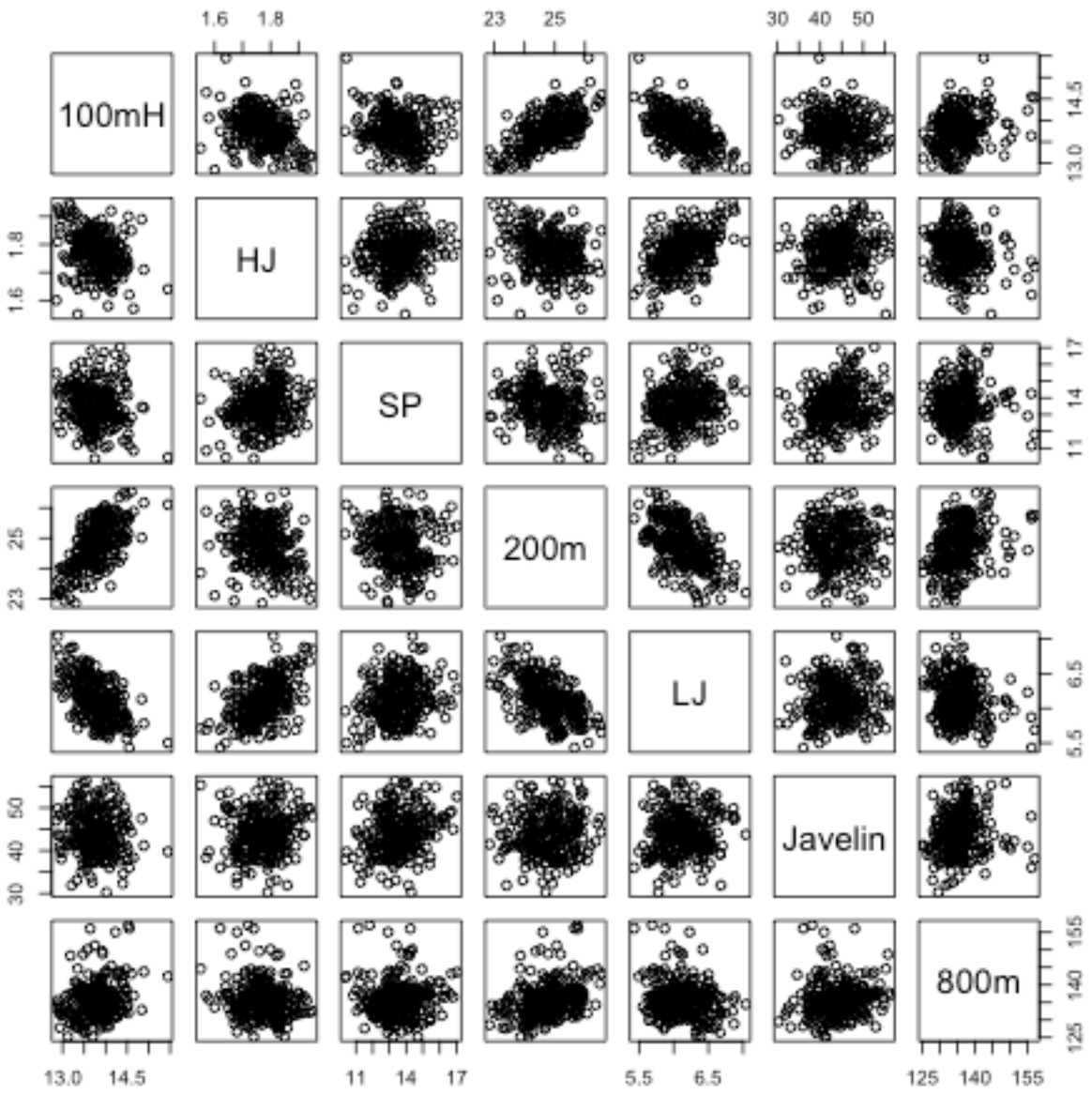
例

- 陸上女子七種競技 (heptathlon)

	100mH	HJ	SP	200m	LJ	Javelin	800m
1	13.32	1.91	13.62	24.49	6.67	48.66	136.09
2	13.02	1.76	13.52	23.98	6.54	43.58	131.48
3	13.70	1.85	13.23	24.13	6.33	40.96	125.23
4	13.54	1.88	12.46	24.20	6.18	47.04	133.24
…							
260	13.88	1.68	11.86	24.94	5.76	41.74	139.97

- 1991～2013 世界陸上 (IAAF World Championships in Athletics)
- IAAF のルールにしたがって点数に換算される。このルールは合理的か？

対散布図



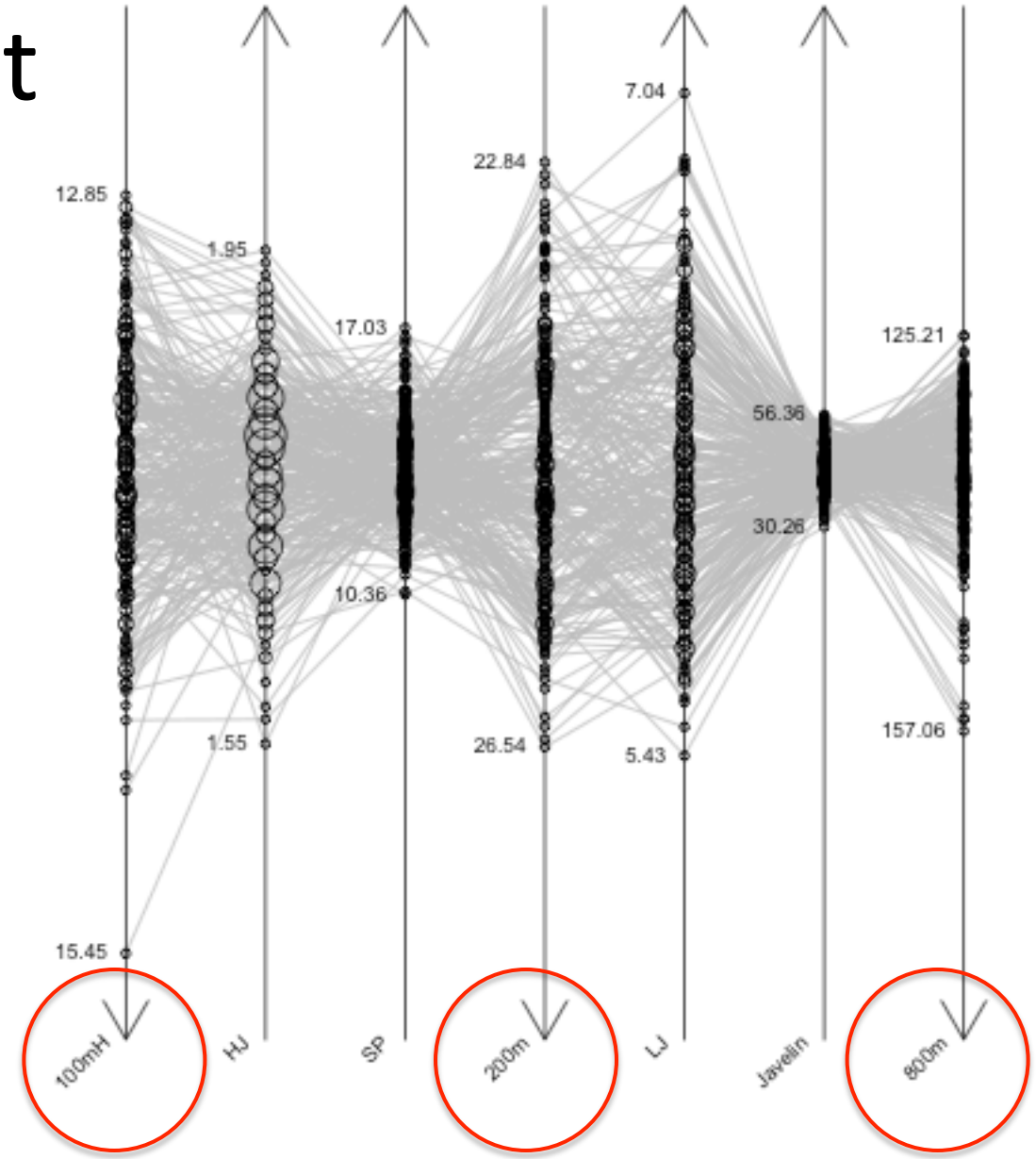
「外れ値」の２名は除外。

Textile Plot

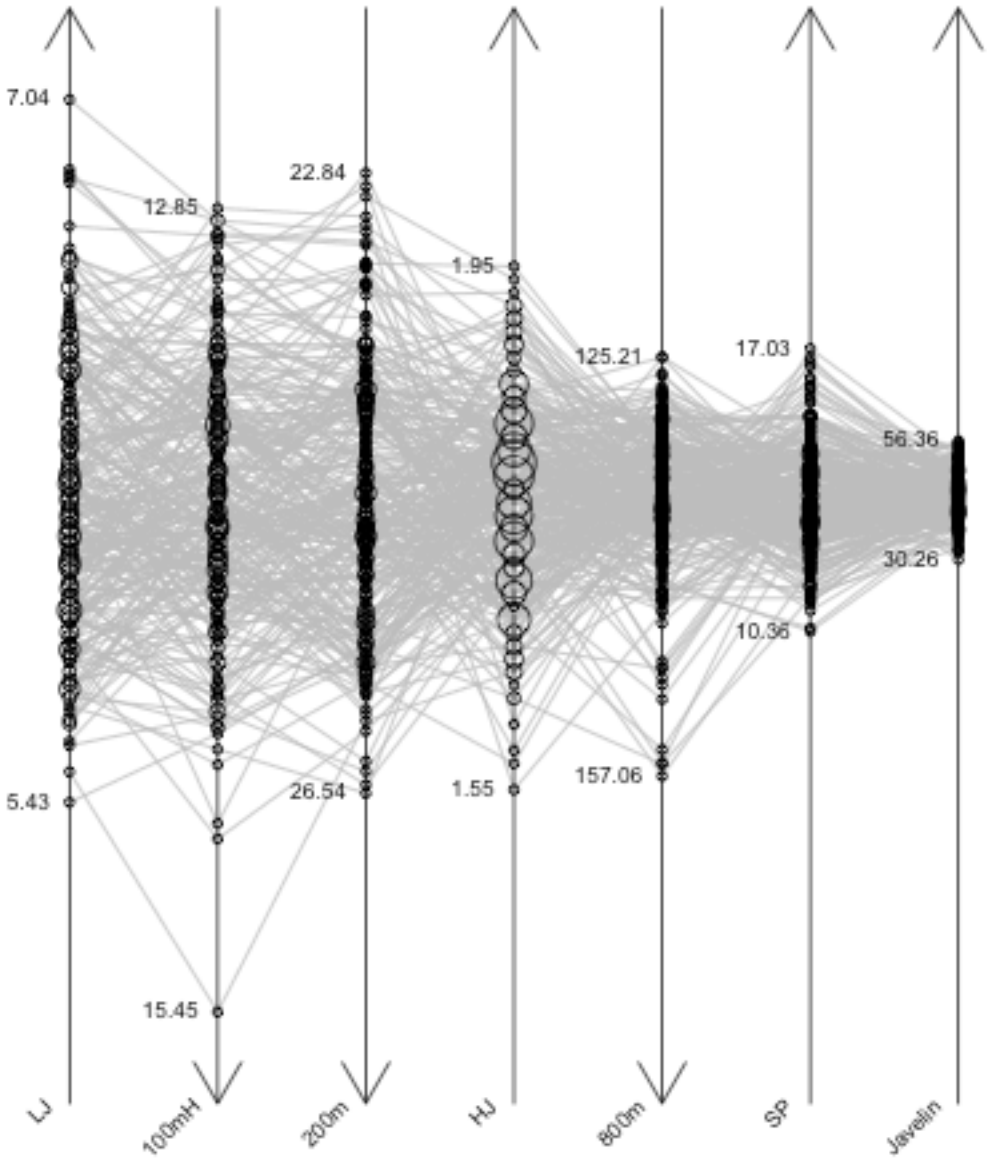
Kumasaka & Shibata (2008)

自作版。
軸の順番は固定。

トラック競技の軸が反転。
合理的。



Textile Plot



参考:
分散の大きい順に
軸を並べ替えた場合

IAAF scoring system

- 七種競技は次の非線形変換で得点化される。

$$y_i = a_i \lfloor d_i x_i - b_i \rfloor^{c_i}$$

	a	b	c	d
100mH	9.23076	26.70	1.835	1
HJ	1.84523	75.00	1.348	100
SP	56.0211	1.50	1.05	1
200m	4.99087	42.50	1.81	1
LJ	0.188807	210.00	1.41	100
Javelin	15.9803	3.80	1.04	1
800m	0.11193	254.00	1.88	1

IAAF scoring system

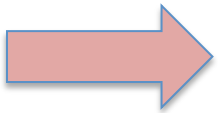
- 変換後、単純和を取ったものが総合得点。

	100mH	HJ	SP	200m	LJ	Javelin	800m	Total
1	1077	1119	769	934	1062	834	877	6672
2	1121	928	762	983	1020	736	943	6493
3	1021	1041	743	968	953	686	1036	6448
4	1044	1080	692	962	905	803	918	6404
…								
260	995	830	652	892	777	701	824	5671

- 以下では変換後の値を元データと見なす。

総合得点はどう決めるべきか

- Z スコアが自然 (e.g. Cox & Dunn 2002 JRSSD)



必ずしも公平とは言えない

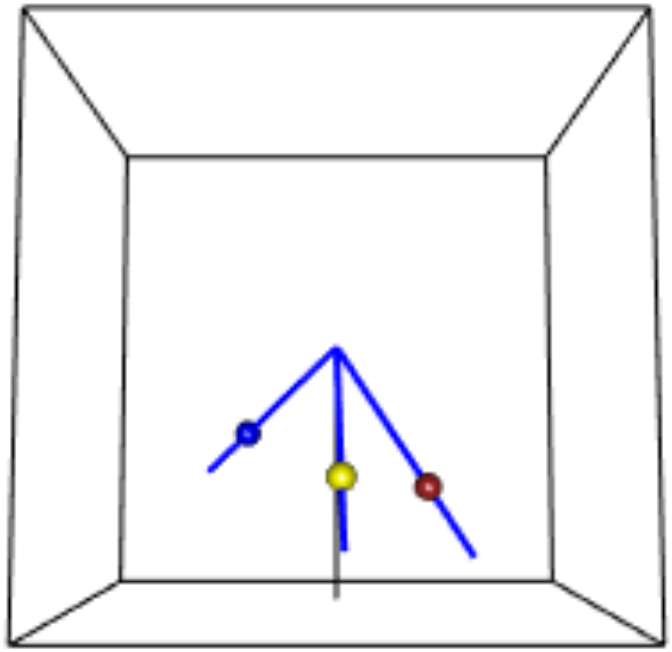
簡単のため3種目で考察

相関係数	100mH	200m	Javelin
100mH	1.00	**0.66**	0.11
200m	---	1.00	-0.04
Javelin	---	---	1.00

この場合、短距離選手が有利になってしまう。

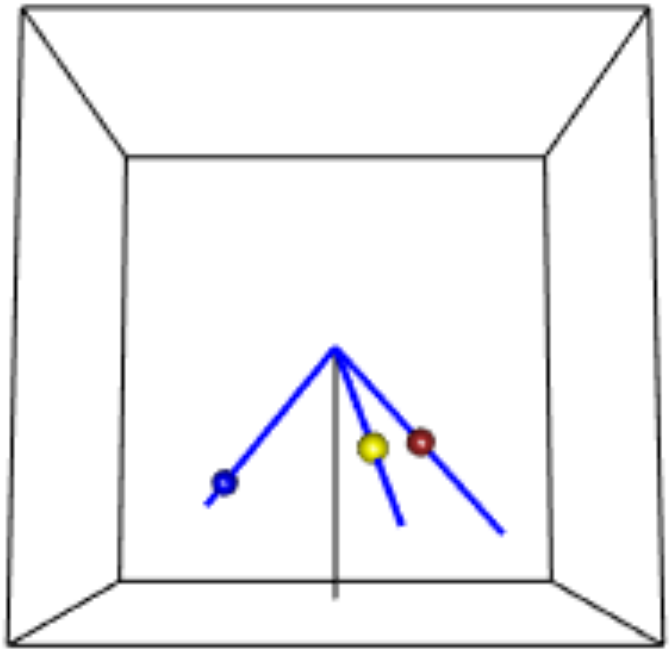
不公平性の図示

- 「やじろべえ」
- 変量空間において、総合得点を鉛直下方向に描いた図。
- 内積が相関係数に対応

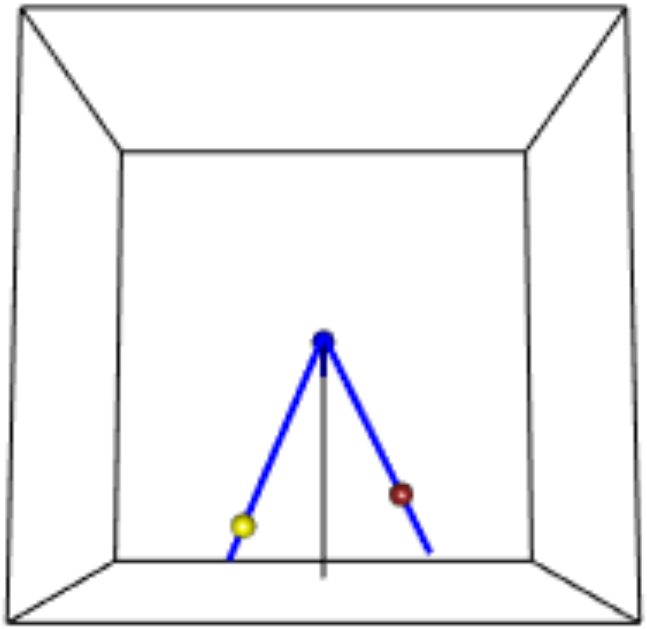


Z スコアの場合

不公平性の図示



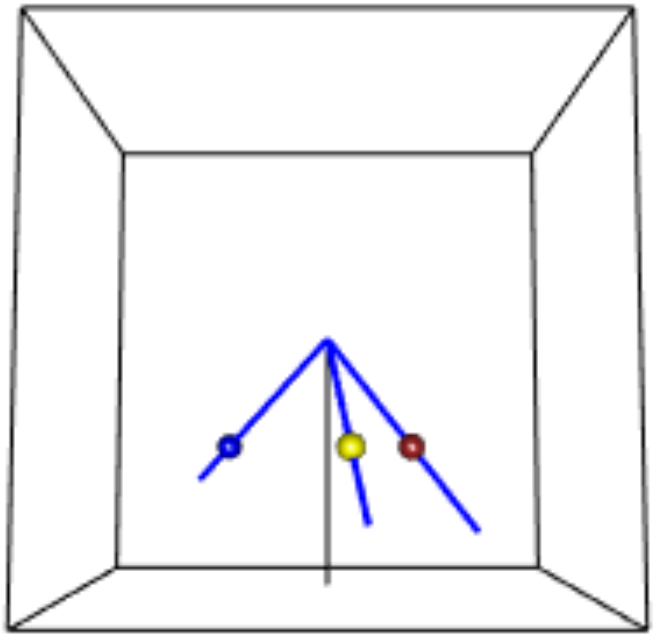
IAAFルールの場合



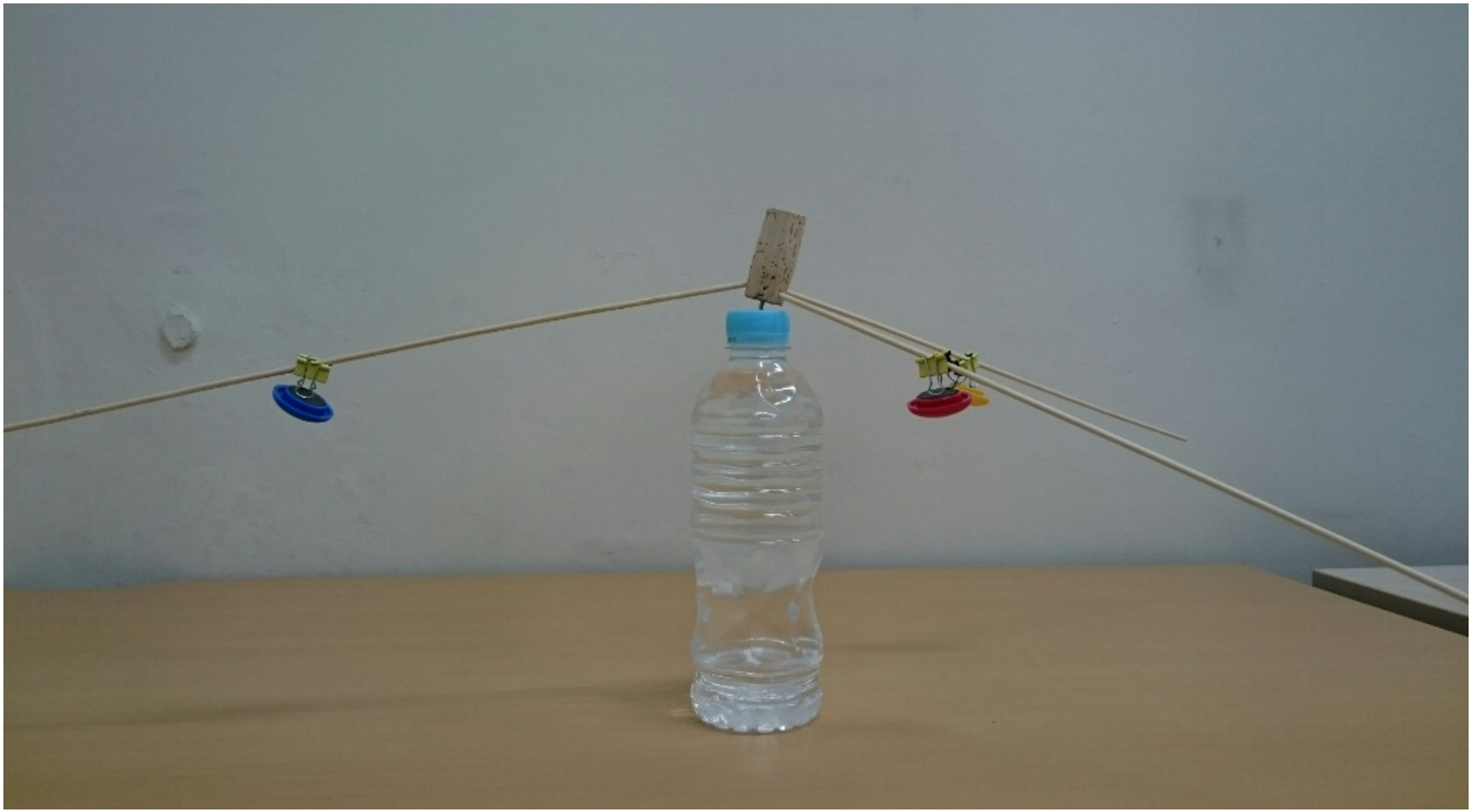
主成分の場合

提案手法

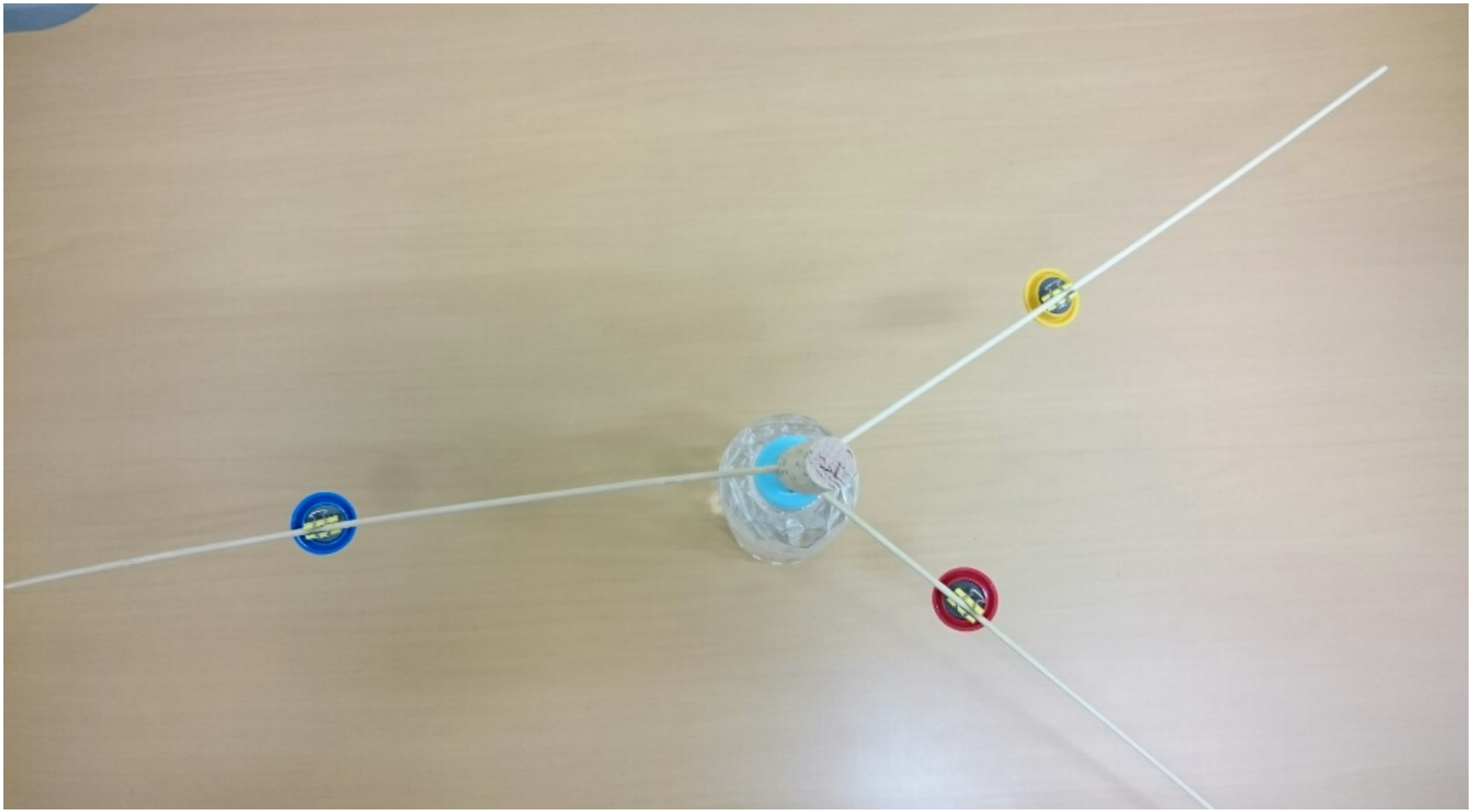
- おもりの位置を調節すると「水平」になる。



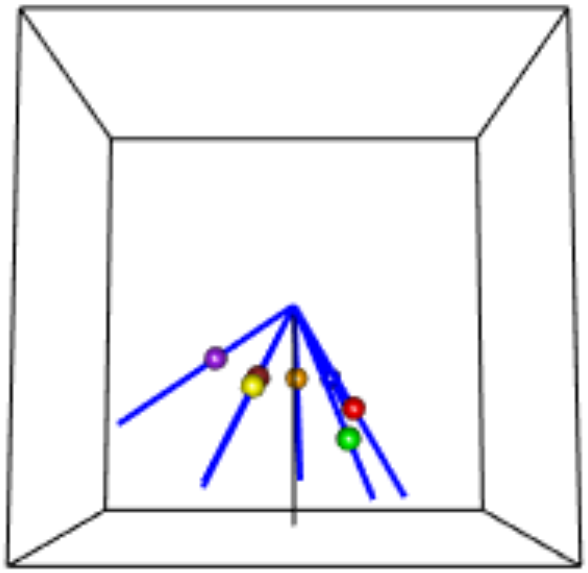
実際作れます



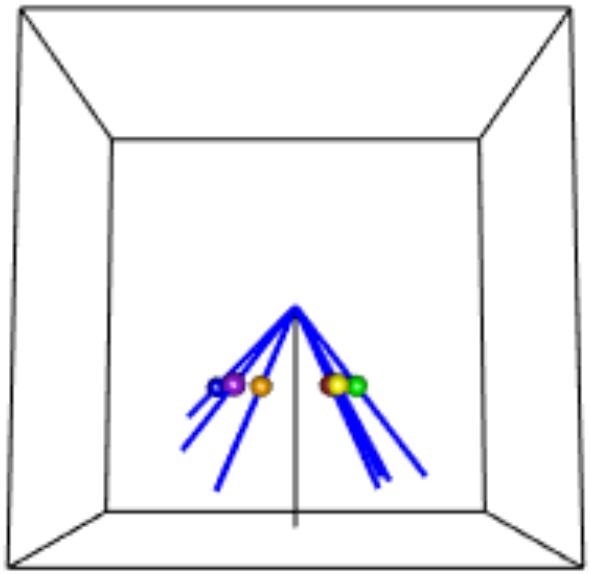
実際作れます



七種全部で図示



IAAFルール



提案手法

- 100mH
- HJ
- SP
- 200m
- LJ
- Javelin
- 800m

7次元空間を3次元に射影している

提案手法

変量 $\mathbf{x}_1,\ldots,\mathbf{x}_p$

重み $w_1,\ldots,w_p$

（七種競技では p = 7）

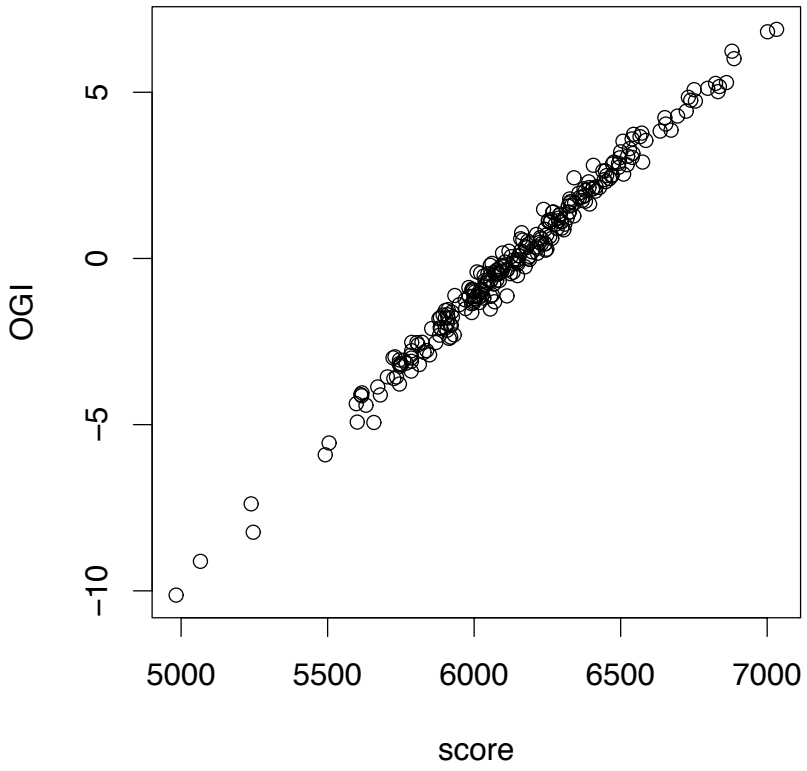
定理 (Marshall & Olkin, 1968) $\exists w_1,\ldots,w_p > 0$

$$\mathbf{g} = \sum_{i=1}^{p} w_i \mathbf{x}_i, \qquad \frac{1}{n}(w_i \mathbf{x}_i)^{\mathrm{T}} \mathbf{g} = 1$$

この $\mathbf{g}$ を客観的総合指数 **OGI** と呼ぶ。

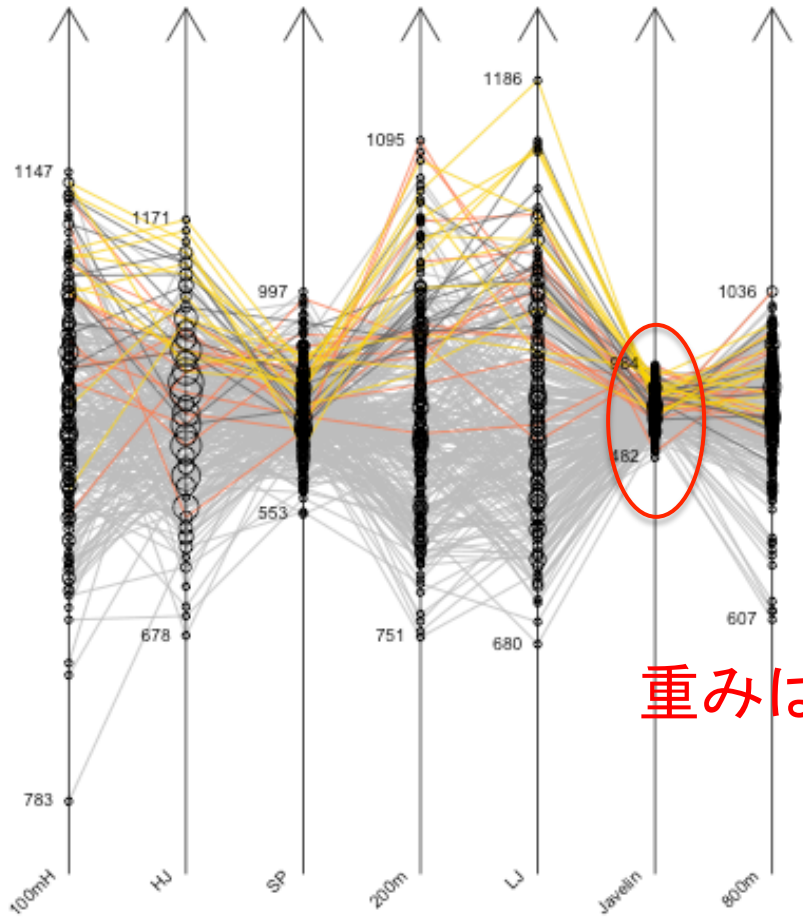
（OGI = objective general index）

七種競技への適用結果

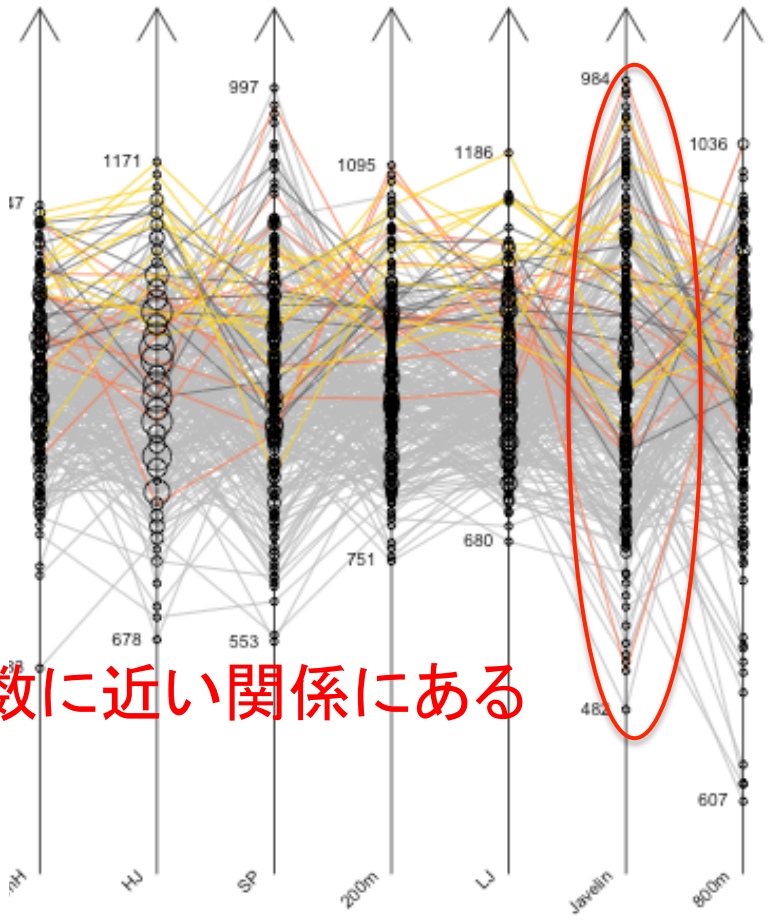


OGI vs. 実際のスコア

TextilePlot との比較



TextilePlot



重みは逆数に近い関係にある

OGI Plot

数理的に比較

$$\mathbf{g} = \sum_{i=1}^{p} w_i \mathbf{x}_i$$

$\mathbf{x}_i$ は標準化済みとする

- Textile Plot

$$\text{Maximize} \quad \mathbf{g}^{\mathrm{T}}\mathbf{g} \quad \text{subject to} \quad \sum_{i=1}^{p} w_i^2 = \text{const.}$$

- OGI

$$\text{Minimize} \quad \mathbf{g}^{\mathrm{T}}\mathbf{g} \quad \text{subject to} \quad \sum_{i=1}^{p} \log w_i = \text{const.}$$

数理的に比較

目的関数と制約を交換して考えると

- Textile Plot

$$\text{Minimize} \quad \sum_{i=1}^{p} w_i^2 \quad \text{subject to} \quad \mathbf{g}^{\mathrm{T}}\mathbf{g} = \text{const.}$$

- OGI

$$\text{Maximize} \quad \sum_{i=1}^{p} \log w_i \quad \text{subject to} \quad \mathbf{g}^{\mathrm{T}}\mathbf{g} = \text{const.}$$

「エントロピー最大化」

まとめ

- 客観的な総合指数 OGI を提案した。
- 女子七種競技への適用例を示した。
- TextilePlot との類似点・相違点を考察した。

- 今後の研究の方向性
  - 景気指標 （共同研究中）
  - 非線形変換を許す場合 （鋭意執筆中）
  - TextilePlot との融合、操作性の向上