SPORTS RANKING: IDENTIFYING CLUSTERS OF EXTREME PREFERENCES

Rosaria Simone, Maria Iannario

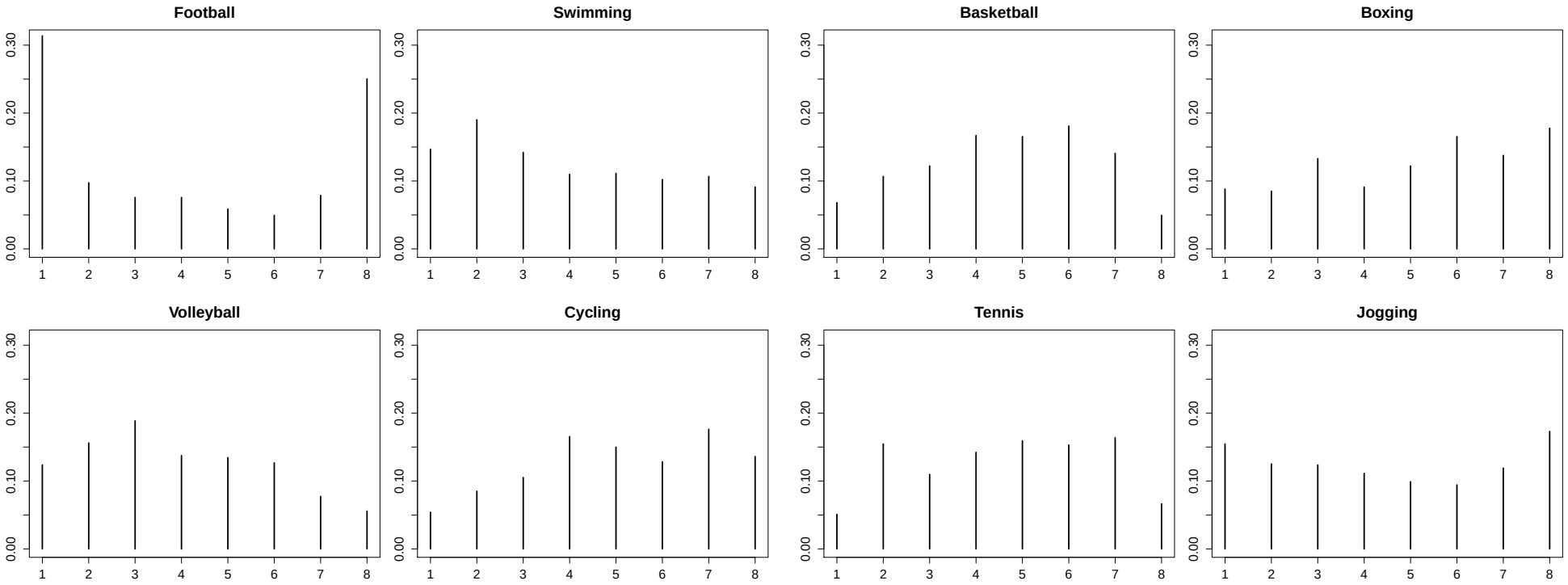
Università degli Studi di Napoli Federico II

November 16th, 2017

WORKSHOP BDSports: BIG DATA ANALYTICS IN SPORTS
Università degli Studi di Brescia
Dipartimento di Economia e Management

Sport preferences

A questionnaire on sport preferences and habits was administered through the web link http://bodai.unibs.it/BDSports/Ricerca2%20-%20DataInn.htm in 2016. Interviewees ($n = 647$) were asked to express their ranking on the 8 most popular sports.



Outline

1. Extreme Preferences
2. Competing models

1 Extreme Preferences

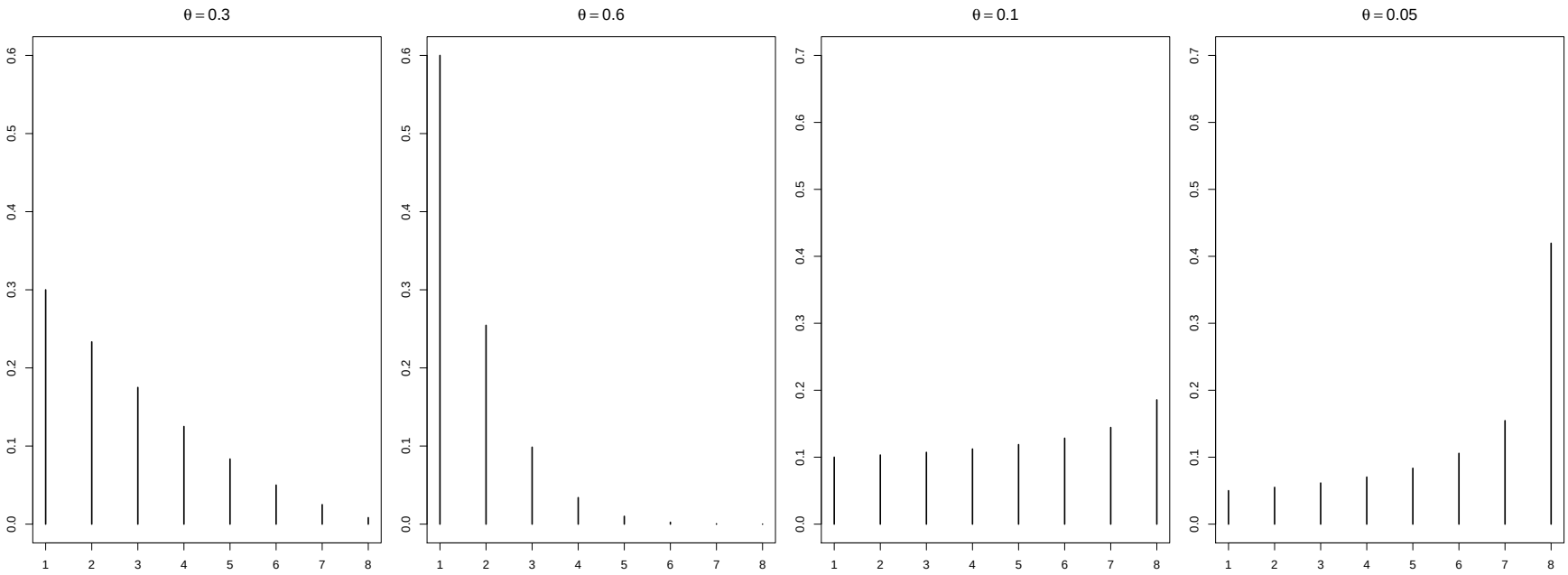
2 Competing models

IHG MODEL

- The Inverse Hypergeometric distribution[1] (IHG) allows to deal with extreme preferences of items (listed as the worst or the best):

$$I_j(\theta_i) = Pr\,(Y = j) = \theta_i(1-\theta_i)^{j-1}\,\frac{m-1}{m}\,\prod_{l=1}^{j}\,\frac{m-l+1}{m-l+\theta_i(l-1)}\,, \quad j = 1,\ldots,m.$$

- with $\theta_i = Pr(Y = 1|\mathbf{x}_i)$ being a direct measure of preference: modal values are constrained to the extreme categories, with $\theta_i > \frac{1}{m}$ ($\theta_i < \frac{1}{m}$) expressing liking (disliking), $logit(\theta_i) = \mathbf{x}_i\,\boldsymbol{\omega}$.



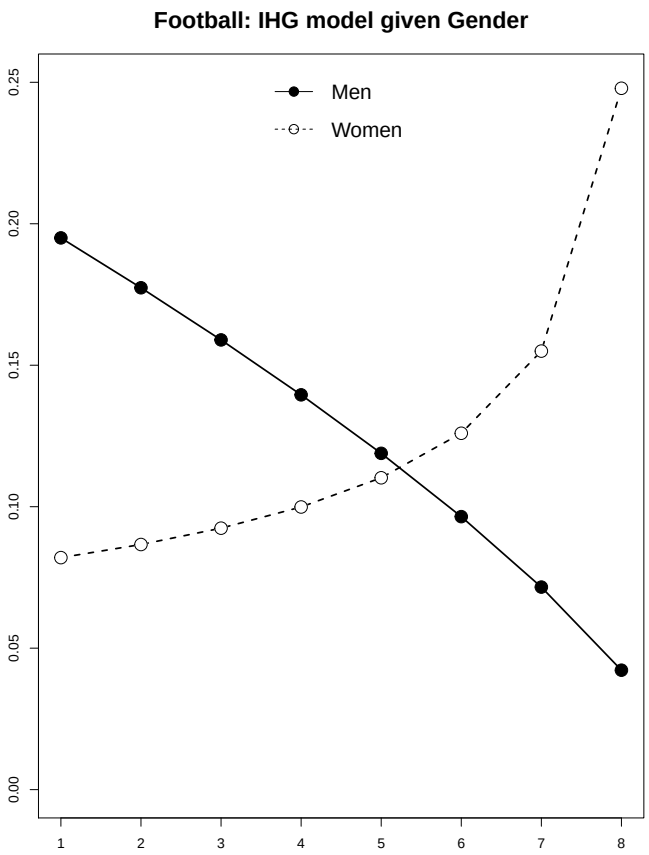
[1] D'Elia, A. (2003), Modelling ranks using the inverse hypergeometric distribution, *Statistical Modelling: an International Journal*, 3, 65–78

Football & gender classification

An IHG model with covariate *Gender* gives the following results ($BIC = 2614.32$):

$$logit(\theta_i) = -1.423 - 0.997\, Gender_i$$

$(\theta | Gender = 0) = 0.195, (\theta | Gender = 1) = 0.082$ [2]



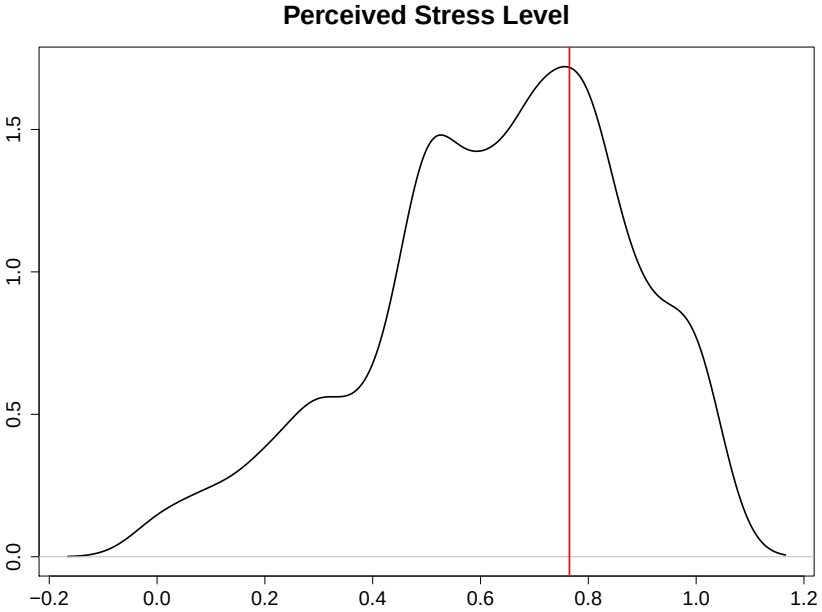
[2] Women ($Gender = 1$) are 60% of the sample!

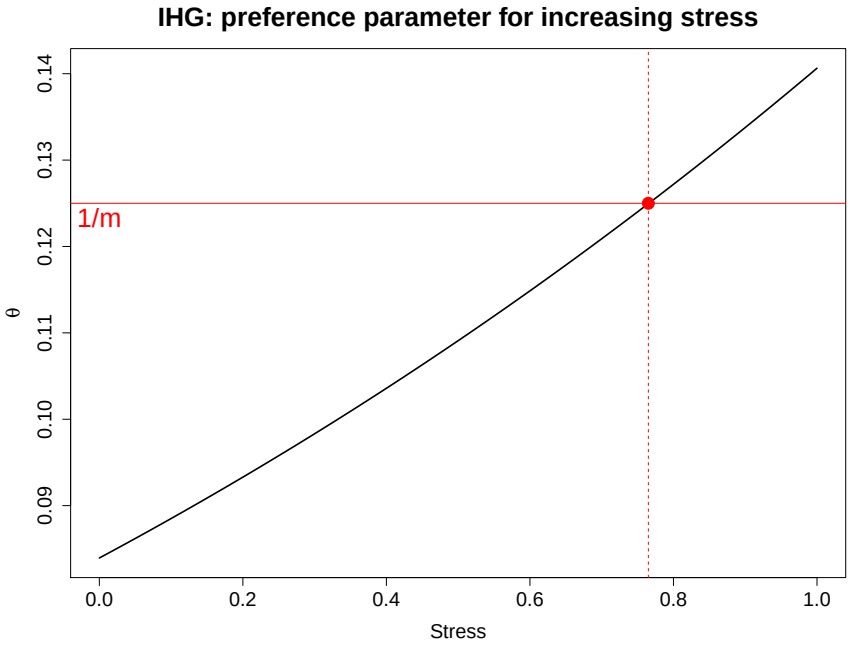
Jogging & Stress Level

- An IHG model with covariate *Stress* (normalized into $(0,1)$) gives the following results ($BIC = 2693.119$):

$$logit(\theta_i) = \underset{(0.14)}{-2.39} + \underset{(0.21)}{0.575}\, Stress$$

- About 68% of respondents expressed a Stress level: $Stress \leq 0.75$





Friedman, E., Berger, B. (1991). Influence of gender, masculinity and femininity on the effectiveness of three stress reduction techniques: Jogging, relaxation response, and group interaction, *Journal of Applied Sport Psychology*, **3**, 61–86.

Schnohr, P., Kristensen, T.S., Prescott, E., Scharling, H. (2005). Stress and life dissatisfaction are inversely associated with jogging and other types of physical activity in leisure time *Scand. J. Med. Sci. Sports*, **15**, 107–112.

Modelling Extreme Preferences

A two-component mixture of IHG distributions (MIHG)[3], say with parameters θ_1 and θ_2:

$$Pr(Y = j|\boldsymbol{\theta}) = \pi\, I_j(\theta_1) + (1-\pi)\, I_j(\theta_2), \quad j = 1, \ldots, m, \tag{1}$$

can be specified to determine groups with opposite preferences.

- For identifiability:

$$0 < \theta_1 < \frac{1}{m} < \theta_2 < 1$$

- One assumes that $\pi > 1 - \pi$ to avoid exchangeability of components; then, the estimation procedure will assign the highest estimated mixing proportion to the first component, with preference parameter θ_1.

[3] Iannario & Simone (2017), Mixture models for analysing ranking data on sport preferences, *IES conference: book of short papers* ISBN: 978-88-8399-107-3.

Modelling Extreme Preferences

A two-component mixture of IHG distributions (MIHG)[3], say with parameters θ_1 and θ_2:

$$Pr(Y = j|\boldsymbol{\theta}) = \pi\, I_j(\theta_1) + (1-\pi)\, I_j(\theta_2), \quad j = 1, \ldots, m, \tag{1}$$

can be specified to determine groups with opposite preferences.

- For identifiability:

$$0 < \theta_1 < \frac{1}{m} < \theta_2 < 1$$

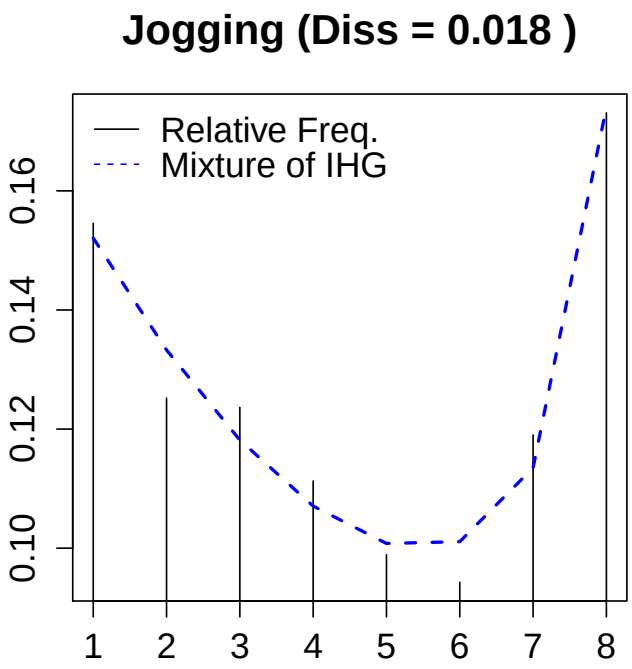
- One assumes that $\pi > 1 - \pi$ to avoid exchangeability of components; then, the estimation procedure will assign the highest estimated mixing proportion to the first component, with preference parameter θ_1.

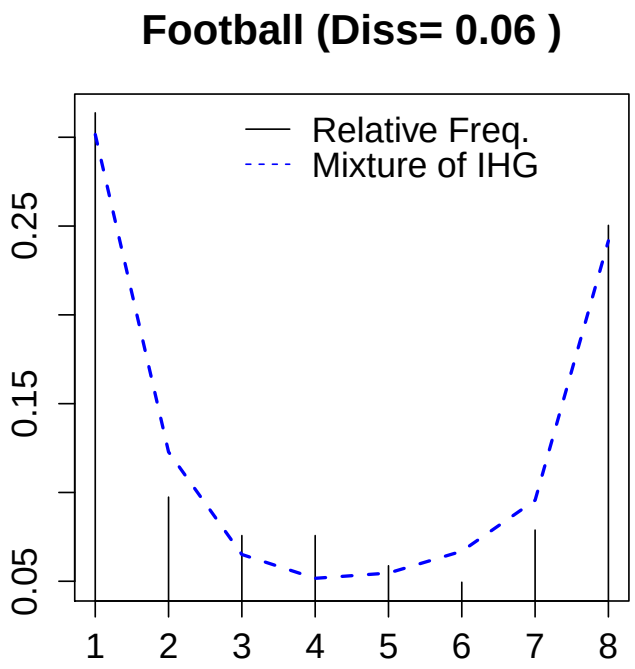
[3] Iannario & Simone (2017), Mixture models for analysing ranking data on sport preferences, *IES conference: book of short papers* ISBN: 978-88-8399-107-3.

Football and Jogging marginals

	$\hat{\pi}$	$\hat{\theta}_1$	$\hat{\theta}_2$	BIC
Football	0.610 *(0.036)*	0.054 *(0.005)*	0.690 *(0.066)*	2422.01
Jogging	0.681 *(0.148)*	0.081 *(0.014)*	0.304 *(0.093)*	2686.10

$$Diss = \frac{1}{2}\sum_{r=1}^{m} \mid f_r - p_r(\hat{\boldsymbol{\theta}}) \mid ; \quad 0 \leq Diss \leq 1$$





1 Extreme Preferences

2 Competing models

Mixture models with uncertainty for ordinal variables

The class of CUB mixture models [4] for rating variables $(Y_1, \ldots, Y_n)$ is grounded on the specification of an *uncertainty* and a *feeling* component:

$$Pr\left(Y_i = j \mid \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\theta}\right) = \pi_i \, b_j\left(\xi_i \mid \boldsymbol{w}_i\right) + \left(1 - \pi_i\right) \frac{1}{m}, \qquad j = 1, \ldots, m$$

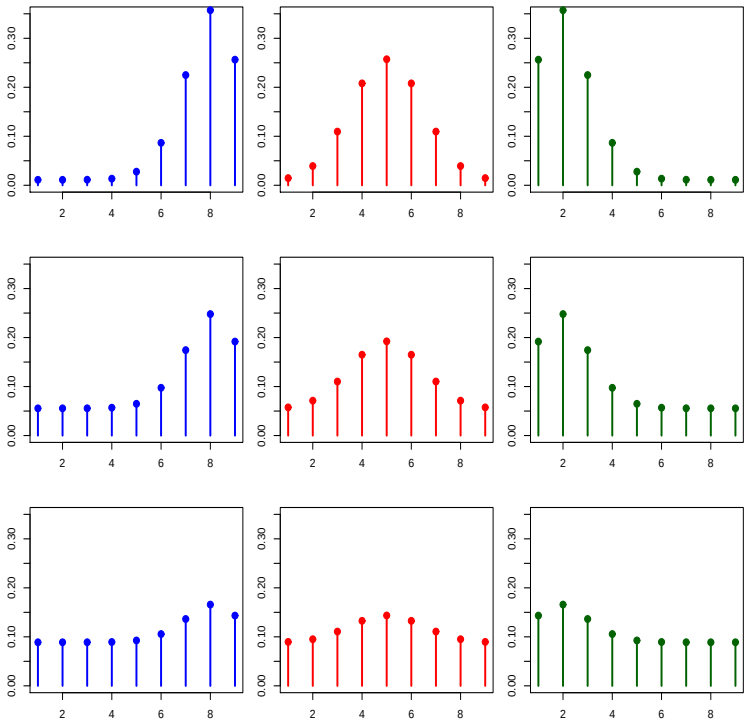
Shifted Binomial:

$$b_j(\xi_i) = \binom{m-1}{j-1} \xi_i^{m-j} (1 - \xi_i)^{j-1}$$

Systematic components:

$$logit\left(\pi_i\right) = \boldsymbol{x}_i \boldsymbol{\beta}$$

$$logit\left(\xi_i\right) = \boldsymbol{w}_i \boldsymbol{\gamma}$$



[4] D'Elia & Piccolo (2005). A mixture model for preferences data analysis, *Computational Statistics & Data Analysis*, 49, 917–937

Inflated categories

GeCUB model [5] If $D_j^{(c)}$ is a degenerate random variable with mass probability concentrated at $j = c$, then the CUB distribution with *shelter effect* at $j = c$ is defined, for $m > 4$, by:

$$Pr(Y_i = j \mid \pi_i, \xi_i, \delta_i) = \delta_i \, D_j^{(c)} + (1 - \delta_i) \left[\pi_i \, b_j(\xi_i) + (1 - \pi_i) \, \frac{1}{m} \right] \tag{2}$$

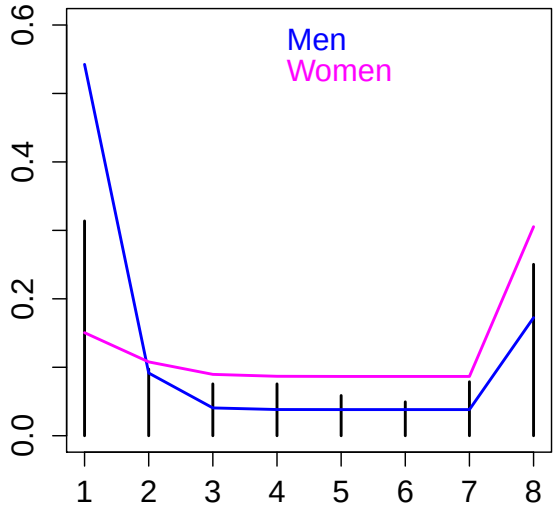
with:

$$logit\,(\delta_i) = \boldsymbol{s}_i \boldsymbol{\omega}$$

GeCUB (shelter at $c = 8$):
$BIC = 2313.58$

	$\hat{\pi}$	$\hat{\xi}$	$\hat{\delta}$	Dissim
Men	0.647	0.985	0.134	0.026
Women	0.113	0.955	0.219	0.031

Football: CUB with shelter at c=8



[5] Iannario & Piccolo (2016), A generalized framework for modelling ordinal data. *Statistical Methods and Applications*, 25, 163–189.

Measurements errors

In general, a person's observed score on a scale is a function of two components:

1. a true score
2. a measurement error:
   - random
   - systematic ⇒ Response Styles: *Consistent and stable tendencies in response behavior which are not explainable by question content or presentation. These are considered to be a source of biased reporting [...]* [6]

ERS **Extreme response style**: is the tendency to select the end points of a rating scale, and it is generally motivated by anxiety, rigidity or poor cognitive structure, intolerance of ambiguity

ARS **Acquiescence response style** ('Yeasaying'): is the tendency to agree with questions, also when in doubt (characteristic of stimulation-seeking extroverts, who tend to accepts statements impulsively)

- **Spike responses** (i.e. midpoint response style): it occurs when some categories receive peculiar attention for some special circumstances, as when people prefer even or odd numbers, or multiples of 5 or 10 (⇒ *Shelter effect*)
- ...

"**Satisficing**" behavior (Simon, 1957):

$$Pr(Y_i = j|\boldsymbol{\theta}_i^\star) = \pi\, b_j(\xi_i) + (1-\pi)\left[(1-\eta_i)\frac{1}{m} + \eta_i\, D_j^{(c)}\right]$$

[6] Cross Cultural Survey Guidelines

Measurements errors

In general, a person's observed score on a scale is a function of two components:

1. a true score
2. a measurement error:
   - random
   - systematic ⇒ Response Styles: *Consistent and stable tendencies in response behavior which are not explainable by question content or presentation. These are considered to be a source of biased reporting [...]* [6]

ERS **Extreme response style**: is the tendency to select the end points of a rating scale, and it is generally motivated by anxiety, rigidity or poor cognitive structure, intolerance of ambiguity

ARS **Acquiescence response style** ('Yeasaying'): is the tendency to agree with questions, also when in doubt (characteristic of stimulation-seeking extroverts, who tend to accepts statements impulsively)

- **Spike responses** (i.e. midpoint response style): it occurs when some categories receive peculiar attention for some special circumstances, as when people prefer even or odd numbers, or multiples of 5 or 10 (⇒ *Shelter effect*)
- ...

If $Pr(S = r)$ is a model for response styles and uncertainty, then:

$$Pr(Y_i = j|\boldsymbol{\theta}_i^\star) = \pi\, b_j(\xi_i) + (1-\pi)\, Pr(S_i = j)$$

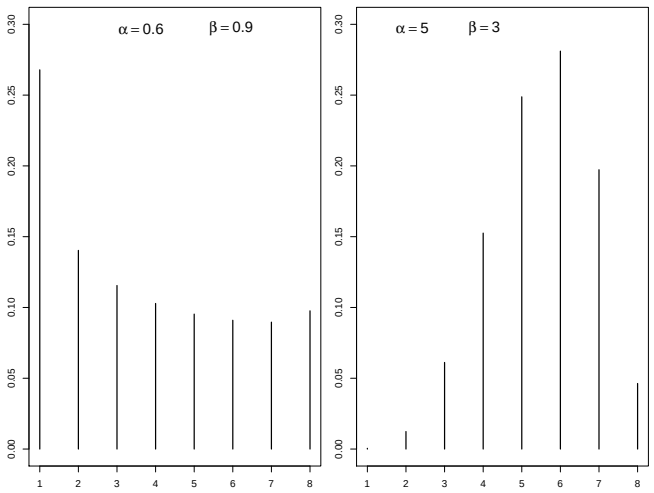
[6] Cross Cultural Survey Guidelines

Discretized Beta Distribution[a]

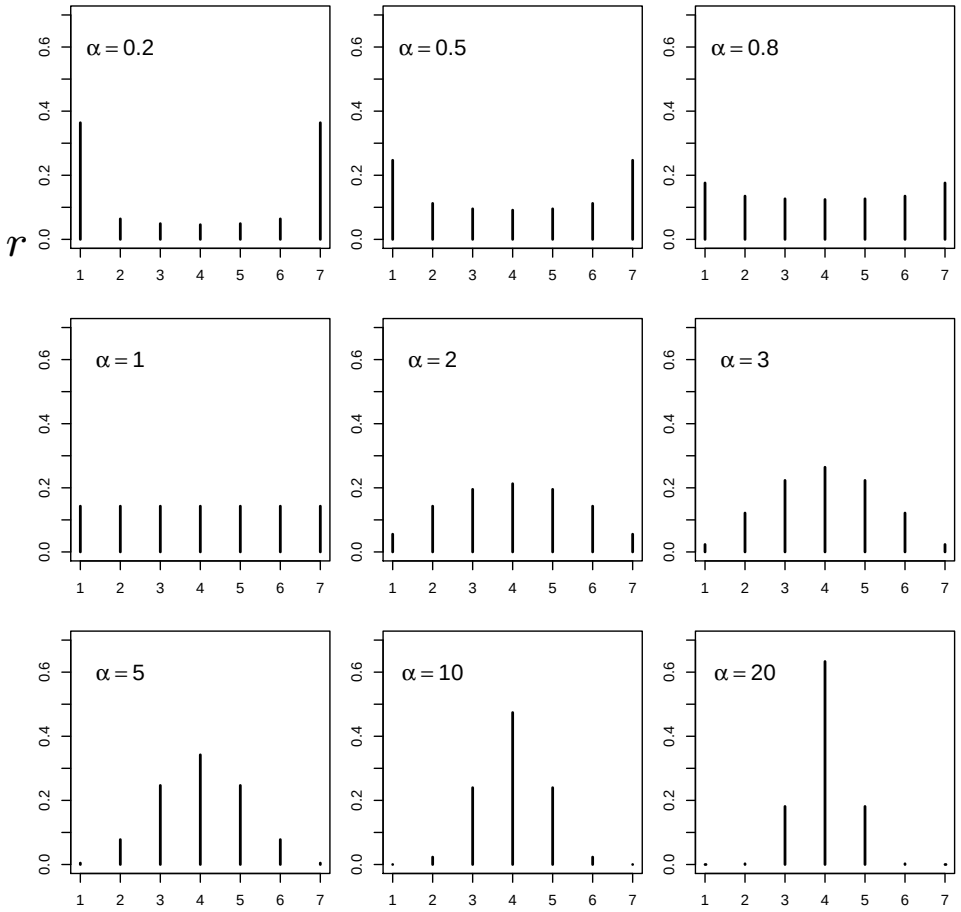
[a] Ursino M. (2014). Ordinal Data: a new model with applications, *PhD Thesis*, `http://porto.polito.it/2535701/`, Politecnico di Torino, Italy.

For $\alpha, \beta > 0$, let $X \sim Beta(\alpha, \beta)$ be a Beta-distributed random variable.
A discrete random variable $D = D(\alpha, \beta)$ over the support $\{1, \ldots, m\}$ has the **discretized Beta distribution** if:

$$Pr(D = r|\alpha, \beta) = Pr\left(\frac{r-1}{m} \leq X \leq \frac{r}{m}\Big|\alpha, \beta\right),$$



If $\alpha = \beta$, D has symmetric distribution:

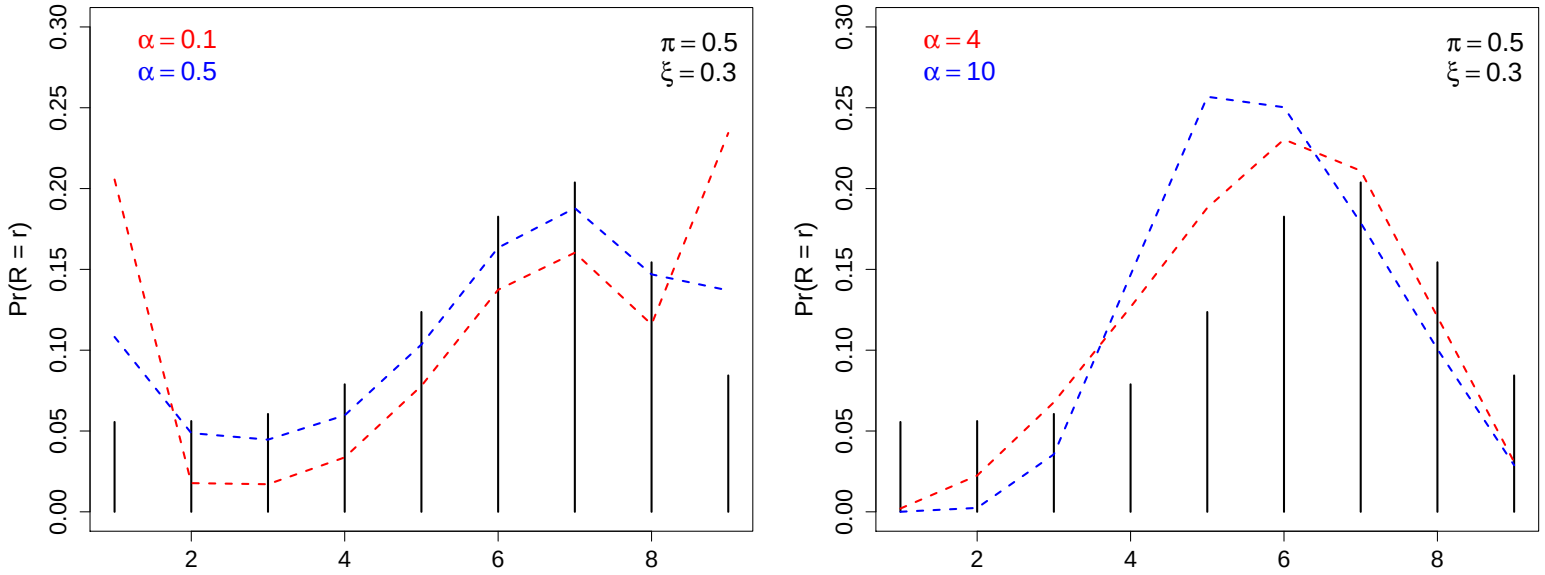


Adjusted Uncertainty

The CAUB model[1], a short for *C*ombination of *A*djusted *U*niform and a shifted *B*inomial, is an extension of the CUB model:

$$Pr(R_i = r|\boldsymbol{\theta}_i) = \pi\, b_r(\xi_i) + (1-\pi)\, Pr(D_i = r|\alpha_i),$$

where D_i follows a discretized Beta distribution $D(\alpha_i)$.



[1] Simone and Tutz, *Modelling Uncertainty and Response Styles in Ordinal Data*, under review

Football ranking and Gender effects

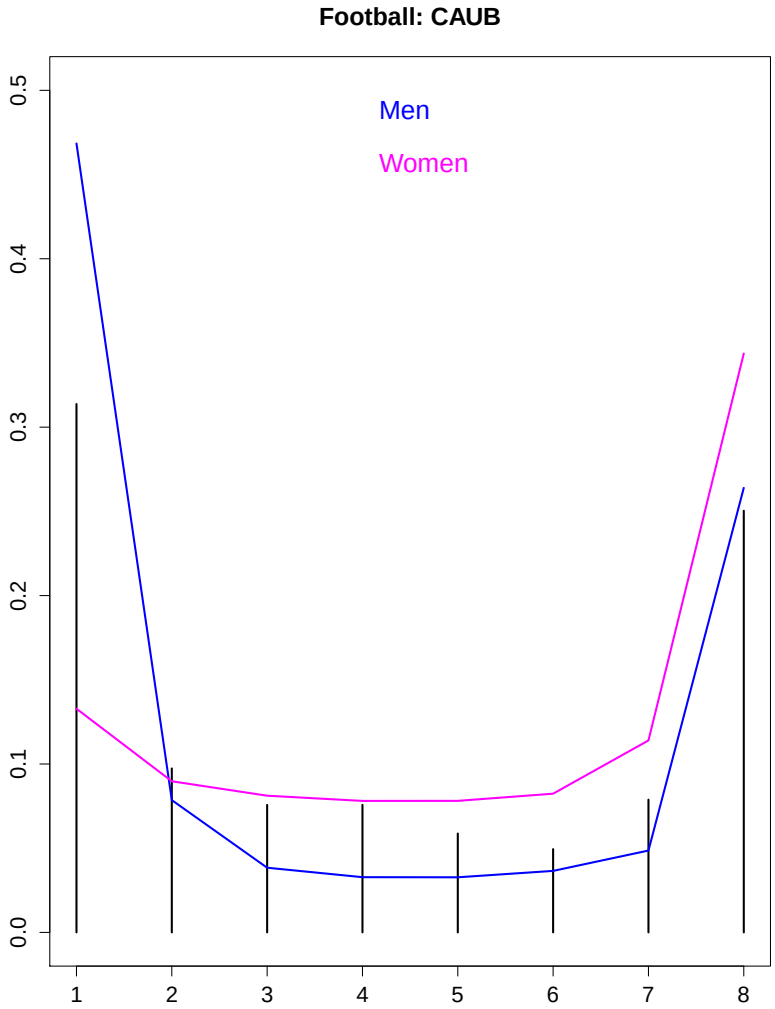
CAUB model fit ($BIC = 2323.89$)

$$\hat{\pi} = \underset{(0.032)}{0.236}$$

$$logit(\hat{\xi}_i) = \underset{(0.850)}{3.867} - \underset{(1.135)}{7.976}\, Gender_i$$

$$log(\hat{\alpha}_i) = \underset{(0.172)}{-1.541} + \underset{(0.197)}{1.213}\, Gender_i$$

	$\hat{\pi}$	$\hat{\xi}$	$\hat{\alpha}$	Dissim
Men	0.236	0.980	0.214	0.11
Women	0.236	0.016	0.720	0.07



Football: Summarizing...

Models without covariates

	BIC	AIC	ΔAIC$^{(*)}$	Diss
MIHG	2422.01	2408.59	13.19	0.06
CAUB	2410.48	2397.06	1.66	0.07
CUB + she(8)	2411.30	2397.88	2.48	0.03
DB(α,β)	2404.35	2395.4		0.03
DB(α,α)	2403.25	2399.5	4.10	0.06

($*$) Burnham & Anderson, 2002

Models with covariate *Gender*

	BIC	AIC	ΔAIC
IHG	2614.32	2605.37	318.62
CAUB	2323.89	2301.53	14.78
GeCUB	2313.58	2286.75	
DB(α,β)			
DB(α,α)	2365.98	2357.04	70.29
POM	2348.26	2312.48	25.73

- The uncertainty specification leads to better fitting performances
- ...

Jogging: Summarizing...

Models without covariates

	BIC	AIC	ΔAIC	Diss
MIHG	2686.10	2672.68	1.19	0.018
CAUB				
CUB + she(8)				
DB(α,β)	2682.34	2673.39	1.89	0.032
DB(α,α)	2675.96	2671.49		0.033

Models with covariate *Stress*

	BIC	AIC	ΔAIC
IHG	2693.12	2684.17	
CAUB			
GeCUB			
DB(α,β)			
DB(α,α)			
POM	2711.74	2675.96	30.35

- The MIHG is insightful to detect latent classes and is of simple specification
- ...

Jogging ranking and stress effects

CAUB model fit (BIC = 2681.39):

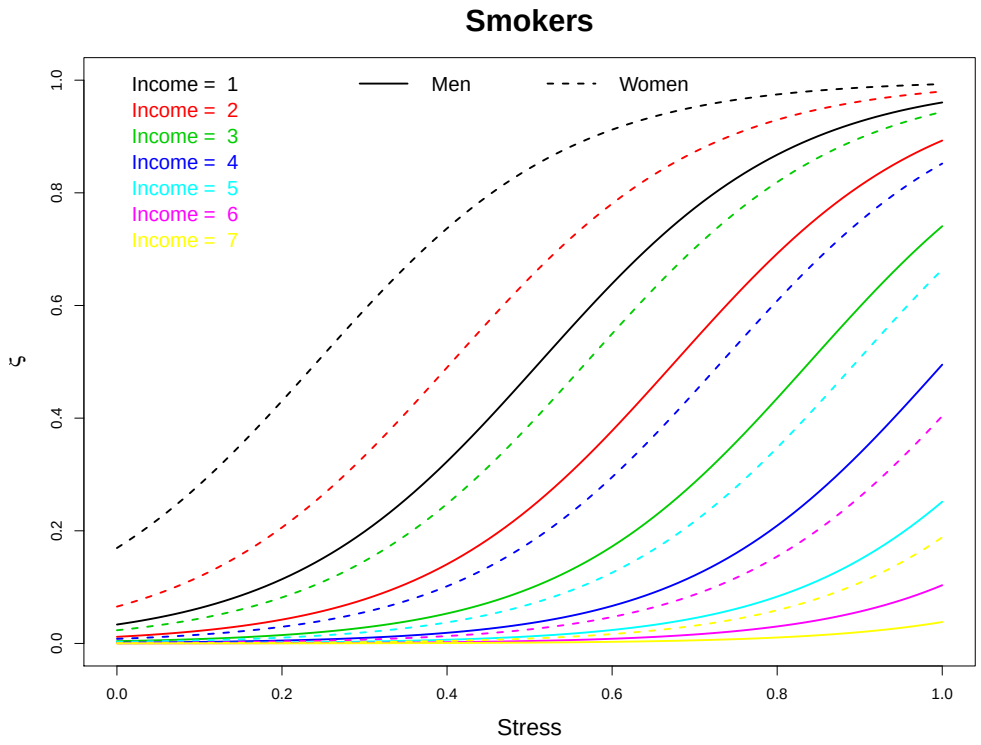
$$\hat{\pi} = \underset{(0.028)}{0.137}$$

$$logit(\hat{\xi}_i) = \underset{1.55}{-6.07} - \underset{(0.308)}{1.07}\, Income_i + \underset{(1.66)}{6.55}\, Stress_i + \underset{(1.07)}{4.77}\, Gender_i + \underset{(0.88)}{2.71}\, Smokers_i$$

$$log(\hat{\alpha}_i) = \underset{0.301}{-1.15} + \underset{(0.05)}{0.182}\, Sport\ \&\ PhysicalWB_i$$

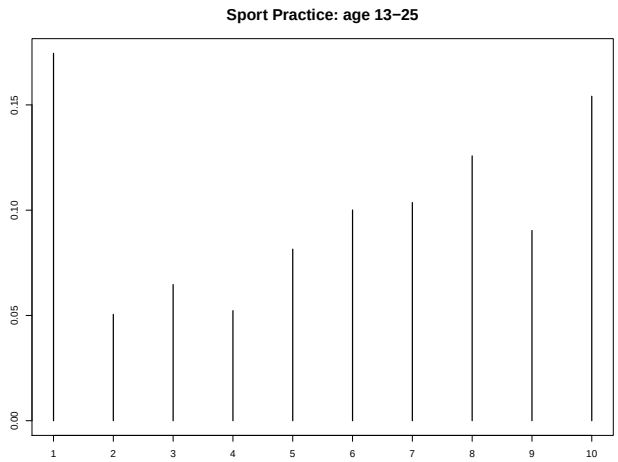
Importance of Sport-Practice in Physical Well-Being

	$\hat{\alpha}$
1	0.317
2	0.380
3	0.456
4	0.547
5	0.656
6	0.787
7	0.944



What's next? Beyond the U-curve...

- An observational study to examine interpersonal relationships and leisure habits of people living in the metropolitan area of Naples, Italy, has been carried out by the Department of Political Sciences, University of Naples Federico II in December 2014
- $n = 2366$ respondents filled the questionnaire where the items describing relational goods, happiness and **engagement in leisure time activities** had to be rated on a scale from 1 to $m = 10$ (1 meaning *"Never, Not at all"* and 10 standing for *"Always, Totally"*).



Forthcoming for BDSports:
Sport participation and Health Behaviour survey

Peretti-Watel, P., Beck, F. Legleye, S. (2002), Beyond the U-curve: the relationship between sport and alcohol, cigarette and cannabis use in adolescents, *Addiction*, **97**(6), 707–716.

Thank you for the attention