Integrating Rasch and Compositional modeling for the analysis of social survey data

Antonio Calcagnì
University of Padova, Italy

52nd Scientific Meeting of the Italian Statistical Society

June 17, 2024

Introduction

Rating data capture human attitudes and opinions but may hold more information than typically conveyed.

Traditional questionnaires capture the final responses, missing insights into individuals' **decision-making processes**, which could reveal variations in hesitancy and uncertainty.

Introduction

Rating data capture human attitudes and opinions but may hold more information than typically conveyed.

Traditional questionnaires capture the final responses, missing insights into individuals' **decision-making processes**, which could reveal variations in hesitancy and uncertainty.

Due to the **interplay** of cognitive, affective and contextual factors in the **process of answering** multiple choice tasks, rating data can disclose more information if appropriately handled.

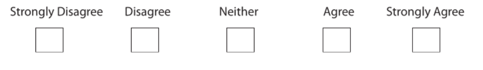
Introduction

Rating as decision-making process

To illustrate this point, consider the question

I am satisfied with my current work

alongside a five-point scale:



Introduction

Rating as decision-making process

To answer,

$\Rightarrow$ the rater retrieves long-term memory information about events and beliefs about their job

Introduction

Rating as decision-making process

To answer

⇒ the rater retrieves long-term memory information about events, beliefs about their job

⇒ these information activate affective components, which influence positively or negatively the **opinion formation** (for example, a recent promotion may enhance the chance for answering the item positively)

Introduction

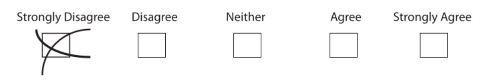
Rating as decision-making process

To answer

⇒ the rater retrieves long-term memory information about events, beliefs about their job

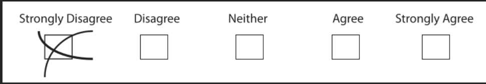
⇒ these information activate affective components which influence positively or negatively the opinion formation (for example, a recent promotion may enhance the chance for answering the item positively)

⇒ cognitive and affective information are integrated to activate the decision making stage and provide the **final response**



Introduction

Rating as decision-making process



Due to conflicting demands in these stages, **decision uncertainty** at certain levels can arise and influence the rating decision.

As a result, the final response does not tell the whole story.

Introduction

Goal: describe a method to extract valuable information from survey responses and analyze them consistently.

Introduction

Goal: describe a method to extract valuable information from survey responses and analyze them consistently.

$\Rightarrow$ via Item Response Theory: Binary tree-based Rasch model [1]

Introduction

Goal: describe a method to extract valuable information from survey responses and analyze them consistently.

⇒ via Item Response Theory: Binary tree-based Rasch model [1]

⇒ via Compositional Data Analysis: Dirichlet regression [2, 4]

Methods

Rasch-IRTree data representation

The **Rasch-tree model** is a member of the IRTrees family [1] and provides a statistical representation of rating responses using **conditional binary trees**.

Methods

Rasch-IRTree data representation

To describe how it works, consider again the previous example:

- I am satisfied with my current work -

| Strongly disagree | — | Disagree | — | 0 | — | Agree | — | Strongly agree |

Methods

Rasch-IRTree data representation

To describe how it works, consider again the previous example:

- I am satisfied with my current work -

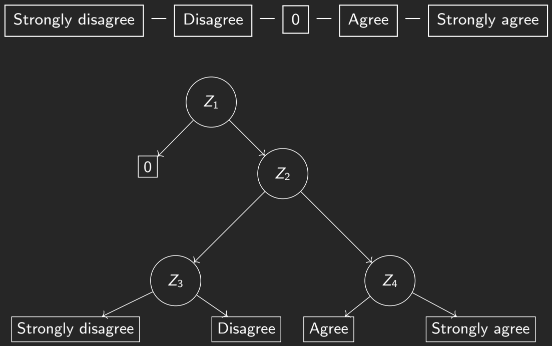
| Strongly disagree | — | Disagree | — | 0 | — | Agree | — | Strongly agree |

Then, each **response option** is thought as being the output of a cognitive sub-process of the entire response process. The sub-processes are modeled as **nodes** of a **binary tree**.

Methods

Rasch-IRTree data representation

An example of 5-point rating scale with the associated binary decision tree.

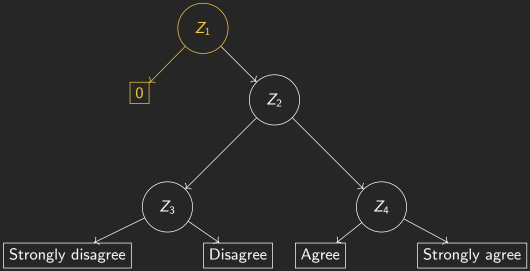


Methods

Rasch-IRTree data representation

In this schema, the rater:

- first decides **whether or not** provide a response ($Z_1 \in \{0, 1\}$)

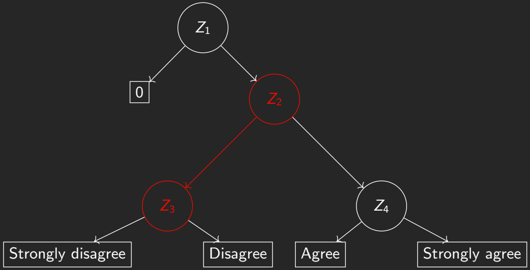


Methods

Rasch-IRTree data representation

In this schema, the rater:

- first decides whether or not provide a response ($Z_1 \in \{0, 1\}$)
- then, for $Z_1 = 1$ he/she decides the **direction** of the response, if negative ($Z_2 = 0$)

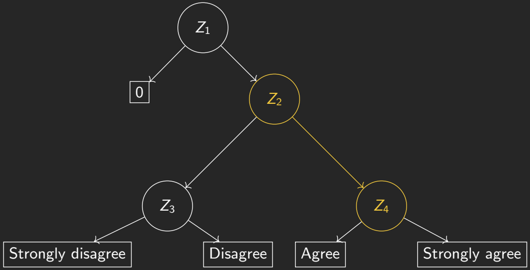


Methods

Rasch-IRTree data representation

In this schema, the rater:

- first decides whether or not provide a response ($Z_1 \in \{0,1\}$)
- then, for $Z_1 = 1$ he/she decides the **direction** of the response, if negative ($Z_2 = 0$) or positive ($Z_2 = 1$)

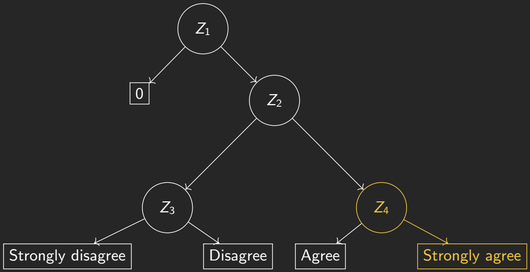


Methods

Rasch-IRTree data representation

In this schema, the rater:

- first decides whether or not provide a response ($Z_1 \in \{0, 1\}$)
- then, for $Z_1 = 1$ he/she decides the direction of the response, if negative ($Z_2 = 0$) or **positive** ($Z_2 = 1$)
- finally, he/she decides the **strength** of the response, e.g. "Strongly agree" ($Z_4 = 1$)



Methods

Rasch-IRTree data representation

The Rasch-tree model is defined by the following equations:
(i-th rater, j-th item, n-th node)

$$Z_{ijn} \sim \mathcal{B}er(\pi_{ijn})$$

$$\pi_{ijn} = \mathbb{P}(Z_{in} = 1; \boldsymbol{\theta}_n) = \frac{\exp(\eta_{in} + \alpha_{jn})}{1 + \exp(\eta_{in} + \alpha_{jn})}$$

$$\eta_{in} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$$

where

$\alpha_{jn} \in \mathbb{R}$: easiness of the **item** being rated

$\eta_{in} \in \mathbb{R}$: **rater**'s latent ability to answer the question

Methods

Rasch-IRTree data representation

The Rasch-tree model is defined by the following equations:
(i-th rater, j-th item, n-th node)

$$Z_{ijn} \sim \mathcal{B}er(\pi_{ijn})$$

$$\pi_{ijn} = \mathbb{P}(Z_{in} = 1; \boldsymbol{\theta}_n) = \frac{\exp(\eta_{in} + \alpha_{jn})}{1 + \exp(\eta_{in} + \alpha_{jn})}$$

$$\eta_{in} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$$

where

$$\mathbb{P}(Y_i = m; \boldsymbol{\theta}_n) = \prod_{n=1}^{N} \mathbb{P}(Z_{in} = d; \boldsymbol{\theta}_n)^d$$

is the probability of the response $Y_i = m$ for the item being rated.

Methods

Rasch-IRTree data representation

The Rasch-tree model is defined by the following equations:
(*i*-th rater, *j*-th item, *n*-th node)

$$
\begin{aligned}
& Z_{ijn} \sim \mathcal{B}er(\pi_{ijn}) \\
& \pi_{ijn} = \mathbb{P}(Z_{in} = 1; \boldsymbol{\theta}_n) = \frac{\exp(\eta_{in} + \alpha_{jn})}{1 + \exp(\eta_{in} + \alpha_{jn})} \\
& \eta_{in} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)
\end{aligned}
$$

The parameters $\boldsymbol{\theta}_n = \{\boldsymbol{\alpha}, \boldsymbol{\Sigma}_\eta\}$ can be estimated via **marginal maximum likelihood** [1].

Methods

Rasch-IRTree data representation

Once $\hat{\alpha}$ and $\widehat{\boldsymbol{\Sigma}}_\eta$ have been recovered conditioned on a sample of data $\mathbf{Y}_{I\times J}$, the estimated **transition probabilities**

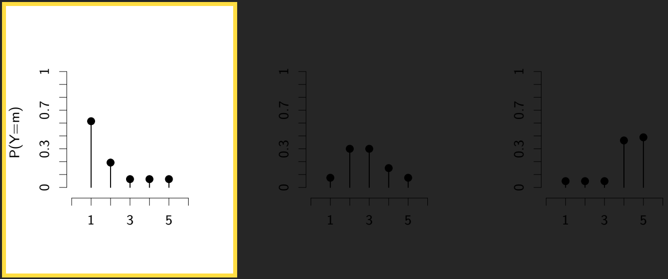
$$\mathcal{U}_i = \Big(\hat{\mathbb{P}}(Y_i = 1), \ldots, \hat{\mathbb{P}}(Y_i = m), \ldots, \hat{\mathbb{P}}(Y_i = M)\Big)$$

provide information about the inner mechanisms of the rater's response process.

Methods

Rasch-IRTree data representation

$$\mathcal{U}_i = \left(\hat{\mathbb{P}}(Y_i = 1), \ldots, \hat{\mathbb{P}}(Y_i = m), \ldots, \hat{\mathbb{P}}(Y_i = M)\right)$$

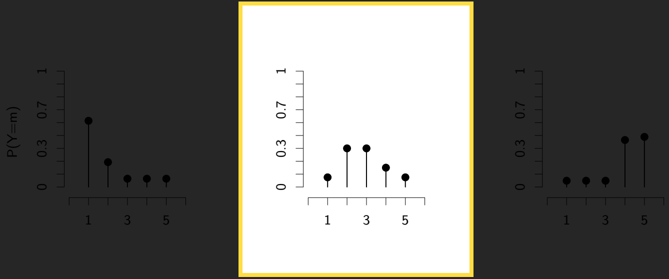


Response process with lower degree of decision uncertainty
(i.e., the response $Y = 1$ is more certain than the remaining ones).

Methods

Rasch-IRTree data representation

$$\mathcal{U}_i = \left(\hat{\mathbb{P}}(Y_i = 1), \ldots, \hat{\mathbb{P}}(Y_i = m), \ldots, \hat{\mathbb{P}}(Y_i = M)\right)$$

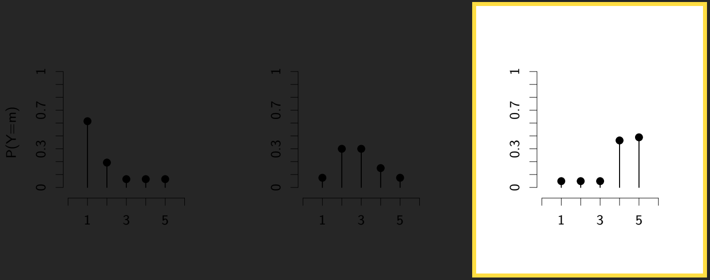


Response process with higher degree of decision uncertainty (i.e., both $Y \in \{2, 3\}$ responses are probable).

Methods

Rasch-IRTree data representation

$$\mathcal{U}_i = \Big(\hat{\mathbb{P}}(Y_i = 1), \ldots, \hat{\mathbb{P}}(Y_i = m), \ldots, \hat{\mathbb{P}}(Y_i = M)\Big)$$



Response process with a certain degree of decision uncertainty (i.e., both $Y \in \{4, 5\}$ responses are probable).

Methods

Rasch-IRTree data representation

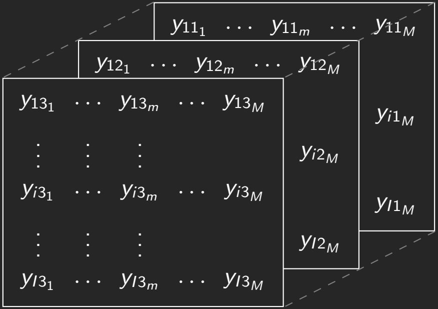
The IRTree procedure can be seen as a form of scale quantification, which outputs a set of M probability masses for each respondent i and survey item j.

$$\mathbf{Y}_{I\times J} \rightarrow \boxed{\text{Rasch-tree}} \rightarrow \widetilde{\mathbf{Y}}_{I\times J\times M}$$

Methods

Rasch-IRTree data representation

$$\mathbf{Y}_{n\times J} \rightarrow \boxed{\text{Rasch-tree}} \rightarrow \widetilde{\mathbf{Y}}_{n\times J\times M}$$



$$y_{ij} \in \{1,\ldots,M\}$$

$$\tilde{\mathbf{y}}_{ij} \in \left\{\mathbf{y} \in \mathbb{R}_+^M : \mathbf{1}^T\mathbf{y} = 1\right\}$$

$$\bar{\mathbf{y}}_{ij} \in [0,M] \subset \mathbb{R} \quad \text{(expectation)}$$

$\Rightarrow$ quantified response

Methods

Dirichlet compositional regression

Let $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_i, \ldots, \widetilde{\mathbf{y}}_n)$ be a collection of independent random compositions with

$$\tilde{\mathbf{y}}_i \in \mathbb{S}^M, \quad \mathbb{S}^M = \left\{ (y_1, \ldots, y_m, \ldots, y_M) \in \mathbb{R}_+^M : \mathbf{1}^T \mathbf{y} = 1 \right\}$$

Let $\mathbf{X}_{n \times J}$ be a matrix of observed variables (e.g., covariates).

The Dirichlet linear model with fixed dispersion ϕ is:

$$\tilde{\mathbf{y}}_i \sim \mathcal{D}(y; \boldsymbol{\mu}_i \phi), \quad g(\boldsymbol{\mu}_i) = \mathbf{x}_i \boldsymbol{\beta}$$

$$\boldsymbol{\mu}_i = \left(\frac{1}{\sum_{m=1}^M \exp \mathbf{x}_i \beta_m}, \frac{\exp \mathbf{x}_i \boldsymbol{\beta}_2}{\sum_{m=1}^M \exp \mathbf{x}_i \beta_m}, \ldots, \frac{\exp \mathbf{x}_i \boldsymbol{\beta}_M}{\sum_{m=1}^M \exp \mathbf{x}_i \beta_m} \right)$$

with $\boldsymbol{\beta}_1 = \mathbf{0}_J$ being the reference level. The parameters estimation is performed via maximum likelihood [4].

Case study

Data description

Aim: Investigate the predictors of anxiety in watching the war in Ukraine [3].

Sample: $n = 796$ respondents from Canada, Germany, and Finland

- 68% female, mean age 24.4 years
- 85% did not have relatives or friends involved in the war

Variables:

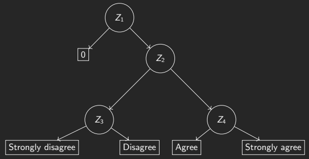
- Response: `anxiety` measured using a six-item questionnaire (with $M = 5$)
- Predictors: `gender` and `depression` (total score from a eight-item questionnaire)

Case study

Data analysis and results

On the first 50% of the dataset ($n = 398$):

- The simple linear decision tree used for modeling `anxiety`
- A random-effect Binomial linear model used to estimate $\hat{\boldsymbol{\alpha}}_{6\times(M-1)}$ for each of the six items and $\hat{\boldsymbol{\Sigma}}_{\eta_{(M-1)\times(M-1)}}$

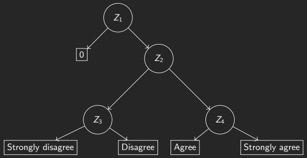


Case study

Data analysis and results

On the second 50% of the dataset ($n = 398$):

- $\hat{\mathbb{P}}(Y_{ij} = 1), \ldots, \hat{\mathbb{P}}(Y_{ij} = M)$ computed for all the items and respondents



Case study

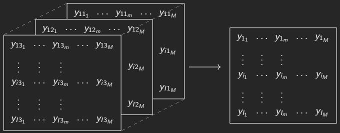
Data analysis and results

On the second 50% of the dataset ($n = 398$):

- $\hat{\mathbb{P}}(Y_{ij} = 1), \ldots, \hat{\mathbb{P}}(Y_{ij} = M)$ computed for all the items and respondents (i.e., $\tilde{\mathbf{Y}}_{I \times J \times M}$)

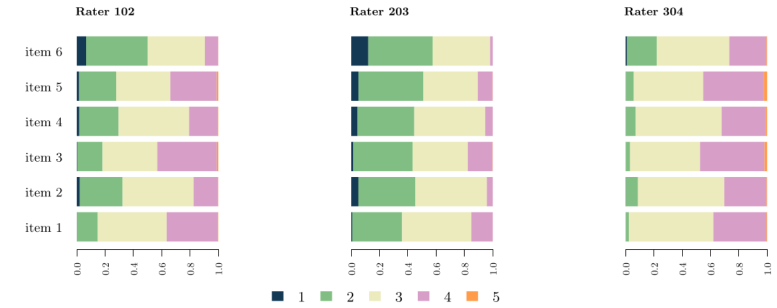
- According to the Aitchinson's geometry, the **compositional total score** of `anxiety` was computed using the perturbation average (i.e., $\tilde{\mathbf{Y}}_{I \times M}$):

$$\tilde{\mathbf{y}}_i = \frac{1}{J} \odot \bigoplus_{j=1}^{J} \mathbf{y}_{ij}$$



Case study

Data analysis and results



Compositional responses for a selection of respondents

Case study

Data analysis and results

On the second 50% of the dataset ($n = 398$):

- Dirichlet linear model with logit (μ) and log (ϕ) links
- μ included `depression`, `gender`, and `depression`$\times$`gender`

Case study

Data analysis and results

On the second 50% of the dataset ($n = 398$):

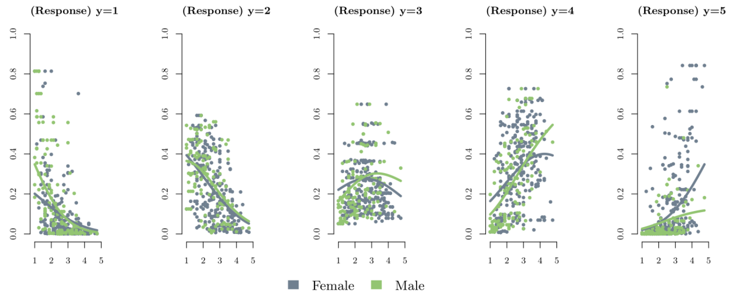
- Dirichlet linear model with logit (μ) and log (ϕ) links
- μ included `depression`, `gender`, and `depression`$\times$`gender`
- $Y = 1$: *"Not at all"* reference level for μ

	$Y=2$			$Y=3$		
	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z
β_0	0.57	0.25	2.28	-0.51	0.24	-2.08
β_{depres}	0.10	0.09	1.11	0.60	0.09	6.74
$\beta_{\text{gender:Male}}$	-1.04	0.37	-2.84	-1.36	0.37	-3.71
$\beta_{\text{depres}\times\text{gender}}$	0.41	0.15	2.69	0.51	0.15	3.41

	$Y=4$			$Y=5$		
	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z
β_0	-1.08	0.24	-4.54	-3.70	0.29	-12.87
β_{depres}	0.87	0.09	10.13	1.40	0.10	13.86
$\beta_{\text{gender:Male}}$	-1.64	0.36	-4.49	-0.27	0.43	-0.64
$\beta_{\text{depres}\times\text{gender}}$	0.57	0.15	3.85	-0.02	0.17	-0.11
ϕ	1.81	0.03	53.03			

Case study

Data analysis and results

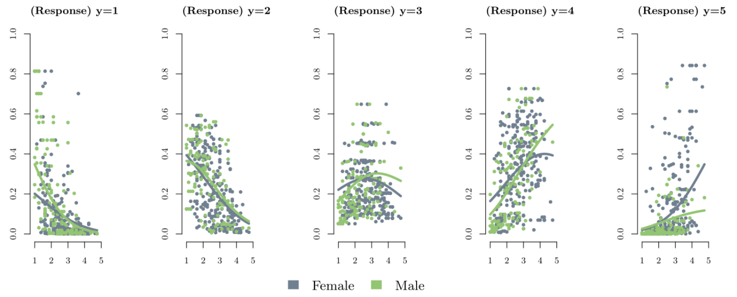


As depression increases, respondents are more likely to self-report substantial anxiety compared to the baseline *"Not at all"*.

The odds of experiencing extreme anxiety ($Y = 5$) increase by a factor of $\exp(\beta_{\text{depres}}) = 4.05$ compared to having no anxiety at all.

Case study

Data analysis and results



The interaction depress×gender reveals that, compared to females with no anxiety, males are approximately 1.70 times more likely to experience moderate ($Y = 3$) or high ($Y = 4$) levels of anxiety when experiencing depression.

Conclusions

- The proposed scale quantification adopts a model-based approach using the IRTree model as a formal representation of the rater's response process.

Conclusions

- The proposed scale quantification adopts a model-based approach using the IRTree model as a formal representation of the rater's response process.

  Cons: It should be valid from a cognitive/theoretical point-of-view.

Conclusions

□ The proposed scale quantification adopts a model-based approach using the IRTree model as a formal representation of the rater's response process.

Cons: It should be valid from a cognitive/theoretical point-of-view.

□ The IRtree statistical model acts as a (kind of) denoiser of the observed data.

Conclusions

- ☐ The proposed scale quantification adopts a model-based approach using the IRTree model as a formal representation of the rater's response process.

  Cons: It should be valid from a cognitive/theoretical point-of-view.

- ☐ The IRtree statistical model acts as a (kind of) denoiser of the observed data.

  Cons: It requires large dataset to estimate α and Σ appropriately.

Conclusions

- The proposed scale quantification adopts a model-based approach using the IRTree model as a formal representation of the rater's response process.

  Cons: It should be valid from a cognitive/theoretical point-of-view.

- The IRtree statistical model acts as a (kind of) denoiser of the observed data.

  Cons: It requires large dataset to estimate α and Σ appropriately.

- It paves the way for analysing data using more informative statistical techniques (i.e., COmpositional Data Analysis).

Conclusions

- The proposed scale quantification adopts a model-based approach using the IRTree model as a formal representation of the rater's response process.

  Cons: It should be valid from a cognitive/theoretical point-of-view.

- The IRtree statistical model acts as a (kind of) denoiser of the observed data.

  Cons: It requires large dataset to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ appropriately.

- It paves the way for analysing data using more informative statistical techniques (i.e., COmpositional Data Analysis).

  Cons: Sometimes, it lacks a clear and immediate interpretation of the results.

[1] Boeck, P. D., and Partchev, I.
IRTrees: Tree-based item response models of the GLMM family.
J. Stat. Soft. 48 (2012).

[2] Filzmoser, P., Hron, K., and Templ, M.
Applied compositional data analysis.
Cham: Springer (2018).

[3] Greenglass, E., Begic, P., Buchwald, P., Karkkola, P., and Hintsa, T.
Anxiety and watching the war in ukraine.
International journal of psychology (2023).

[4] Gueorguieva, R., Rosenheck, R., and Zelterman, D.
Dirichlet component regression and its applications to psychiatric data.
Computational statistics & data analysis 52, 12 (2008), 5344–5355.

antonio.calcagni@unipd.it

https://unipd.link/acalcagni