Discussão dos textos

Employing digital footprints in psychometrics
Education Data Science- Past, Present, Future

Prof. Dr. Ricardo Primi

USF
UNIVERSIDADE SÃO FRANCISCO

Modern Psychometrics

The Science of Psychological Assessment

Fourth Edition

John Rust, Michal Kosinski, and David Stillwell

ROUTLEDGE

8 Employing digital footprints in psychometrics

Introduction

Decades of research and practical applications have shown that well-made tests and self-reported questionnaires can be reliable, practical, and accurate. They have been successfully applied across diverse contexts, ranging from recruitment and high-stakes educational assessments to clinical diagnosis. New methodologies such as computer adaptive testing (see Chapter 5) and item generators (see Chapter 9) are becoming more widespread, further driving the quality of such assessments.

At the same time, however, there are major flaws inherent to self-reported questions and test items. First is their temporal character and low ecological validity: assessments offer only a brief window into respondents' opinions and performance, and they are often administered in an artificial environment, such as in an assessment center and under time pressure. During the brief interaction with a questionnaire or test, a respondent may be affected by factors such as the testing environment, stress, fatigue, or even the weather. Consequently, their scores reflect not only the traits being measured but also these external factors, decreasing the measurement's validity and reliability.

Second, traditional assessments are limited to capturing respondents' explicit, conscious, and motivated opinions and behaviors. Consequently, they are vulnerable to cheating and misrepresentation, particularly when much depends on the scores, such as in the context of recruitment or entrance exams. Misrepresentation is often unconscious, driven by a wide range of unconscious cognitive biases. Availability bias, for instance, leads to overestimating the frequency of thoughts or behaviors that are easily accessible in one's memory. For example, after weeks spent preparing for a job interview, job candidates are likely to underestimate how social they normally are. Another common bias, the reference-group bias, describes the difficulty of comparing one's trait levels with the average in a general population. Instead, we tend to compare ourselves with those around us, or a reference group. An extraverted actor, for instance, might genuinely believe themselves to be introverted if they are surrounded by even more extraverted peers. Consequently, even widely used and well-validated assessments are often relatively poor predictors of many basic real-life outcomes such as performance at work, well-being, or physical activity.

How can we solve these and other limitations of traditional assessments? One approach would be to replace tests and questionnaires—narrow snapshots of respondents' self-reported behaviors—with long-term observations of actual behaviors, preferences, and performance in the natural environment. One could follow the respondents around for, let us say, a full year, meticulously recording all the times when they expressed

Síntese

Limitações de instrumentos de auto relato

Alternativa: "*long-term observations of actual behaviors, preferences, and performance in the natural environment*"

- *Our ongoing migration to the digital environment opened up a myriad of ways in which our behaviors, preferences, and performance can be recorded in an unobtrusive, cheap, and convenient way.*

- *web-browsing logs, records of transactions from online and off-line marketplaces, photos and videos, GPS location logs, media playlists, voice and video call logs, language used in tweets or emails, and much more.*

Digital footprints

Tipos de dados

Síntese

Aplicações

- substitutos de medidas tradicionais, novos contextos e novas medidas (sistemas de recomendação), predição, estudo do comportamento humano, dar suporte a medidas existentes

Vantagens e desafios

- validade ecológica, detalhamento e longitude, menor controle da situação de testagem, velocidade e nao-intrusão, privacidade, ausência de anonimato, viés, enriquecimento de avaliaçào de construtor psicológicos

Desenvolvendo medidas

Dados

- feedback, representatividade, N

Análises

- Matriz de dados
- Matriz gigante e esparsa
- Redução de dimensionalidade (SVD, LDA)

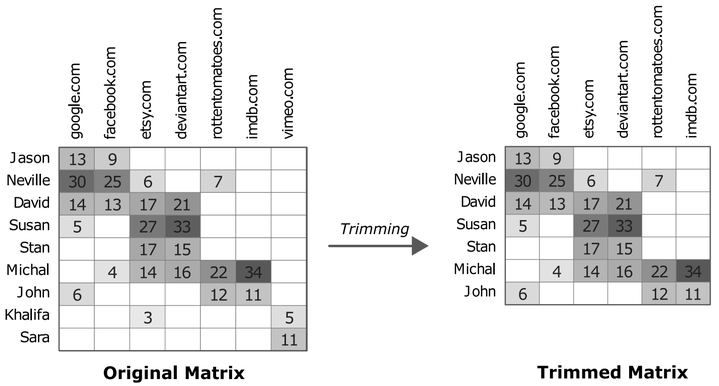

Figure 8.1 A hypothetical respondent-footprint matrix representing the frequencies of website visits and its trimmed version (see text for details). Cells represent the number of times a given respondent visited a given website. Shading based on the frequency was added to enhance readability. Zeros were removed for clarity.

Popular languages for statistical programming (e.g., Python, R, and MATLAB) provide off-the-shelf functions that allow for reducing matrix dimensionality using SVD. To preserve computational resources, make sure to use a sparse SVD function, or a function that can take a sparse matrix as an input without converting it into a nonsparse format. SVD decomposes a matrix into three matrices (**U**, **V**, and **Σ**) exposing its underlying structure. Matrices **U** and **V** contain singular vectors subsuming the patterns present in the original matrix. Diagonal matrix **Σ** contains singular values representing the importance of each of the singular vectors. (A diagonal matrix is a matrix where only the diagonal cells are filled with values.) The product of **U**, **Σ**, and transposed **V** ($\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{T}$) is equal to the original matrix.

The first singular vector subsumes the most prominent pattern in the matrix, and the subsequent vectors represent patterns of decreasing importance. Thus, the dimensionality of the matrix can be reduced by discarding some of the less important singular vectors. The product of the resulting trimmed matrices **U**, **Σ**, and $\mathbf{V}^{T}$ does not represent the original matrix exactly, but provides its approximation.

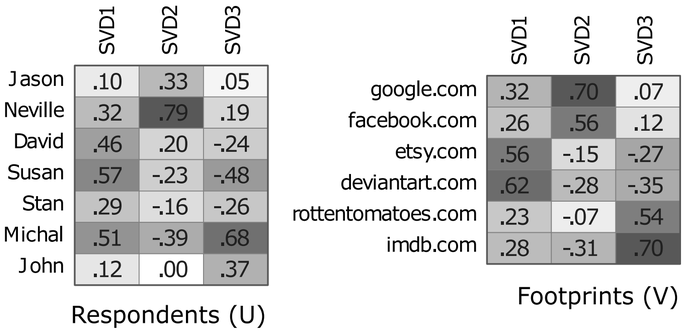

Figure 8.3 Respondents' (matrix V) and websites' (matrix U) scores on three singular vectors extracted from the trimmed respondent-footprint matrix presented in Figure 8.1.

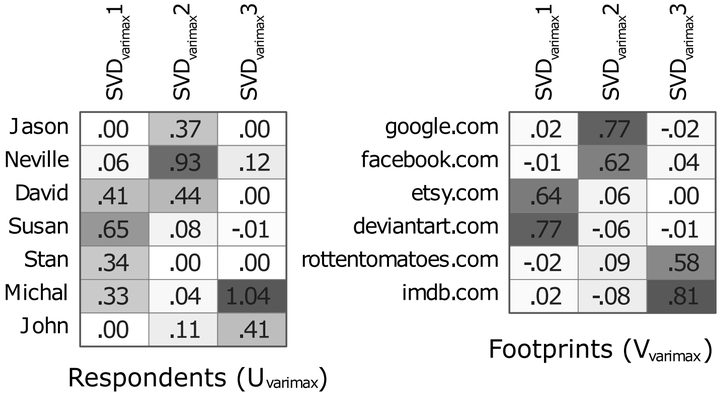

Figure 8.4 Varimax-rotated singular vectors from Figure 8.3.

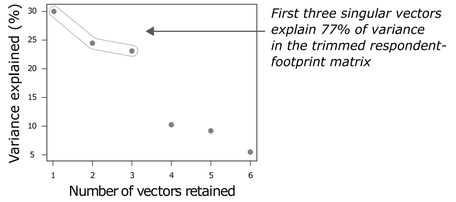

Figure 8.2 Variance explained by consecutive singular vectors in the trimmed respondent-footprint matrix presented in Figure 8.1 (right panel).

Figure 8.6 presents two matrices produced by LDA expressing the associations between clusters and websites (matrix **β**) and clusters and respondents (matrix **γ**). Matrix **β** contains probabilities of a particular website being visited (relatively, within a cluster). For example, take cluster LDA1, which groups art-related websites. When a participant visits a website in this cluster, they will pick Deviantart.com with a probability of .53, Etsy.com with a probability of .46, and Google.com with a probability of .01. Note that the probabilities sum to 1 in each column. This is because we are dealing with mutually exclusive events encompassing all possible outcomes: if a participant visits a given cluster, they must choose one of the websites in the matrix.

Matrix **γ** contains probabilities of a respondent visiting one of the websites in a given cluster. For example, David's probability of visiting websites in cluster LDA1 (Etsy.com and Deviantart.com) equals .6, while his probability of visiting websites in cluster LDA2 (Google.com and Facebook.com) equals .4. Compare this with the respondent-footprint matrix (Figure 8.1) showing that David visited exclusively websites in those clusters, and that he was more likely to visit those belonging to LDA1. In matrix **γ**, the probabilities

	LDA1	LDA2	LDA3
Jason	.00	.99	.00
Neville	.00	1.0	.00
David	.60	.40	.00
Susan	.92	.08	.00
Stan	.99	.00	.00
Michal	.33	.00	.67
John	.00	.28	.72

Respondents (γ)

	LDA1	LDA2	LDA3
google.com	.01	.52	.00
facebook.com	.00	.36	.05
etsy.com	.46	.05	.00
deviantart.com	.53	.00	.00
rottentomatoes.com	.00	.07	.39
imdb.com	.00	.00	.55

Footprints (β)

Figure 8.6 Three LDA topics extracted from the trimmed respondent-footprint matrix presented in Figure 8.1. Matrix γ shows associations between respondents and clusters; matrix β shows the associations between websites and clusters.

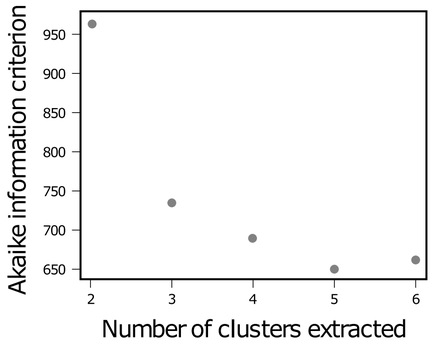

Figure 8.5 AIC and the number of LDA clusters extracted from the trimmed respondent-footprint matrix presented in Figure 8.1. (Note that the minimum number of clusters that can be extracted is k = 2.)

Desenvolvendo modelos preditivos

1. Outcome

2. Dividir amostra em treino / teste

3. Redução de dimensionalidade

4. Escores sujeitos (matriz gama ou U) e treino do modelo

5. Previsão na base teste

6. Validade e precisão

AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–12
DOI: 10.1177/23328584211052055

Education Data Science: Past, Present, Future

Daniel A. McFarland
Saurabh Khanna
Benjamin W. Domingue
Stanford University
Zachary A. Pardos
University of California, Berkeley

This AERA Open special topic concerns the large emerging research area of education data science (EDS). In a narrow sense, EDS applies statistics and computational techniques to educational phenomena and questions. In a broader sense, it is an umbrella for a fleet of new computational techniques being used to identify new forms of data, measures, descriptives, predictions, and experiments in education. Not only are old research questions being analyzed in new ways but also new questions are emerging based on novel data and discoveries from EDS techniques. This overview defines the emerging field of education data science and discusses 12 articles that illustrate an AERA-angle on EDS. Our overview relates a variety of promises EDS poses for the field of education as well as the areas where EDS scholars could successfully focus going forward.

Keywords: *data science, network analysis, natural language processing, machine learning, learning analytics, data mining*

Síntese

“data science” as an alias for computer science and statistics respectively

International Association for Statistical Computing (IASC) in 1977 was established with a “mission to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.”

Data science had risen to occupy a unique disciplinary position on account of (a) being more application oriented as it targets solutions to real-world chal- lenges (Donoho, 2017), (b) coupling quantitative and quali- tative research across disciplines (Dhar, 2013), and (c) being largely focused on digital structured and unstructured data (Silver, 2020).

Learning Analytics, Machine Learning, Artificial Intelligence, Data Science, and Natural Language Processing.

Síntese

“data science” as an alias for computer science and statistics respectively

International Association for Statistical Computing (IASC) in 1977 was established with a “mission to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.”

Data science had risen to occupy a unique disciplinary position on account of (a) being more application oriented as it targets solutions to real-world chal- lenges (Donoho, 2017), (b) coupling quantitative and qualitative research across disciplines (Dhar, 2013), and (c) being largely focused on digital structured and unstructured data (Silver, 2020).

Learning Analytics, Machine Learning, Artificial Intelligence, Data Science, and Natural Language Processing.