Book of Short Papers SIS 2020

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Copyright © 2020

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891910776

Contents

Specialized sessions

Solicited Sessions

Contributed papers and Posters

Models and methods - Categorical, Ordinal, Rank Data 1135

Models and methods – Regression..1206

Ranking extraction in ordinal multi-indicator systems

Costruzione di ranking in sistemi multidimensionali di indicatori ordinali

Marco Fattore and Alberto Arcagni

Abstract In this paper, we present a procedure for scoring and ranking statistical units in ordinal multi-indicator systems, by integrating classical dimensionality reduction tools and novel results in Partial Order Theory. Units are ranked based on "dominance" scores, which depend upon both the structure of the partial order and the joint frequency distribution. Dominance scores are complemented with scores of incomparability among units, so to assess the ranking quality. The procedure is computationally light and is here applied to data about financial literacy in Italy.
Abstract In questo articolo, presentiamo una procedura per la costruzione di ranking di unità statistiche, valutate su sistemi multidimensionali di variabili ordinali. La procedura integra algoritmi di riduzione della dimensionalità e recenti risultati della Teoria degli Ordinamenti Parziali, e ordina le unità in base a punteggi di "dominanza", che dipendono dalla struttura dell'ordinamento parziale e dalla distribuzione di frequenze congiunte. Al punteggio di dominanza è affiancato un punteggio di "incomparabilità", per valutare la qualità del ranking. La procedura è computazionalmente leggera ed è qui esemplificata su dati relativi alla competenza finanziaria in Italia.

Key words: Financial literacy; Ordinal data; Partial order; Poset; Ranking.

Marco Fattore
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1 - 20126 - Milano e-mail: marco.fattore@unimib.it

Alberto Arcagni
Sapienza University of Rome, Piazzale Aldo Moro 5 - 00185 - Roma e-mail: alberto.arcagni@uniroma1.it

1 Introduction

Let $\boldsymbol{v}_1,\dots\boldsymbol{v}_k$ be a system of k ordinal variables, on scales with $m_1,\dots m_k$ degrees, recorded on n statistical units. Each unit u is associated a score profile $\boldsymbol{u} = (u_1,\dots,u_k)$, composed of variable scores, among the $m = m_1\cdot\ldots\cdot m_k$ possible different score configurations. Since the input variables are ordinal, such m score configurations can be partially ordered according to the *product order* rule [1, 2, 4], i.e. given two profiles $\boldsymbol{x}_i = (x_{i1},\dots,x_{ik})$ and $\boldsymbol{x}_j = (x_{j1},\dots,x_{jk})$, we put $\boldsymbol{x}_i < \boldsymbol{x}_j$ if and only if $x_{ih} \le x_{jh} \forall h = 1,\dots,k$ and there exists an index z such that $x_{iz} < x_{jz}$. Many units can share the same profile and so we are naturally led to consider distributions on product orders, as the typical data structure arising from assessing statistical populations against ordinal multi-indicator systems. The classical problem of ranking units based on their multidimensional profiles is thus the problem of reducing the dimensionality of frequency distributions defined on product orders. In the following, we propose an algorithm to solve this issue.

2 The lexicographic dominance and incomparability matrices

As any finite partially ordered set, a product order π over k ordinal variables can be uniquely reconstructed by means of its set of *linear extensions*, i.e. of linear orders obtained from π, by ordering all non-comparable configuration pairs, in all possible ways. As shown in [6], however, being a product order π can also be reconstructed by using the subset of so-called *lexicographic linear extensions* (LLEs). These are linear orders where score configurations are ordered in "alphabetic fashion", with respect to all possible permutations of the k input variables; thus, there are $k!$ LLEs and these suffice to generate the product order associated to $\boldsymbol{v}_1,\dots\boldsymbol{v}_k$. LLEs are the key tool for scoring and ranking units, as they bring information on the relative positions of score configurations, in the product order.

Informally speaking, to rank units we want to associate to each of them a dominance score, reflecting to what degree a unit is placed "higher" than the others, in π. As a first step to this goal, we compute the dominance score P_{ij} between two profiles $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, as the fraction of LLEs in which $\boldsymbol{x}_i$ is placed below $\boldsymbol{x}_j$. Scores P_{ij} are called *lexicographic mutual ranking probabilities* (LMRPs) and can be computed analitically, as shown in the following proposition.

Proposition. Let $\boldsymbol{v}_1,\dots\boldsymbol{v}_k$ be a system of k ordinal variables, on scales with $m_1,\dots m_k$ degrees, and let $\boldsymbol{x}_i = (x_{i1},\dots,x_{ik})$ and $\boldsymbol{x}_j = (x_{j1},\dots,x_{jk})$ be two profiles on it. Let k_1 be the number of indices h such that $x_{ih} < x_{jh}$ and let k_2 be the number of indices h such that $x_{ih} = x_{jh}$. Then:

$$P_{ij} = \frac{k_1}{k}\sum_{s=0}^{k_2}\frac{k_2!}{(k_2-s)!}\frac{(k-s-1)!}{(k-1)!}. \tag{1}$$

Proof. Since the set of LLEs is in bijective correspondence with that of all permutations of variables, we can identify a LLE λ with a permutation $l_1,\ldots,l_k$ of $\{1,\ldots,k\}$. With this notation, the subset G of LLEs such that $\boldsymbol{x}_i < \boldsymbol{x}_j$, can be partitioned as $G = G_0 \cup \ldots \cup G_{k_2}$, where sets G_s and their cardinalities are given by:

1. $G_0 = \{\lambda \,|\, x_{il_1} < x_{jl_1}\} \to |G_0| = k_1 \cdot (k-1)!$
2. $G_1 = \{\lambda \,|\, x_{il_1} = x_{jl_1}, x_{il_2} < x_{jl_2}\} \to |G_1| = k_2 \cdot k_1 \cdot (k-2)!$
3. $G_2 = \{\lambda \,|\, x_{il_1} = x_{jl_1}, x_{il_2} = x_{jl_2}, x_{il_3} < x_{jl_3}\} \to |G_2| = k_2 \cdot (k_2-1) \cdot k_1 \cdot (k-3)!$
4. $\ldots$
5. $G_{k_2} = \{\lambda \,|\, x_{il_1} = x_{jl_1}, \ldots, x_{il_{k_2}} = x_{jl_{k_2}}, x_{il_{k_2+1}} < x_{jl_{k_2+1}}\} \to |G_{k_2}| = k_2 \cdot (k_2-1) \cdot \ldots \cdot 1 \cdot k_1 \cdot (k-k_2-1)! = k_2! k_1 \cdot (k-k_2-1)!$.

Summing up all of the cardinalities and dividing by the total number of LLEs ($k!$), we get

$$
\begin{aligned}
P_{ij} = P(\boldsymbol{x}_i < \boldsymbol{x}_j) &= \frac{|G_0| + \ldots + |G_{k_2}|}{k!} = \frac{k_1}{k}\Big[1 + \frac{k_2}{k-1} + \frac{k_2(k_2-1)}{(k-1)(k-2)} + \\
&+ \frac{k_2(k_2-1)(k_2-2)}{(k-1)(k-2)(k-3)} + \ldots + \frac{k_2!}{(k-1)(k-2)(k-3)\ldots(k-k_2)}\Big] = \\
&= \frac{k_1}{k}\sum_{s=0}^{k_2} \frac{k_2!}{(k_2-s)!}\frac{(k-s-1)!}{(k-1)!}.
\end{aligned}
$$

q.e.d.

LMRPs are used as entries of the *lexicographic dominance matrix* P, which is later used to derive the final dominance scores associated to units. Making all profiles comparable, by means of dominance scores, is unavoidably forcing and stretches the data structure, since it destroys the incomparabilities existing in π. In order to keep control on this kind of "distorsion", we complement matrix P with a *lexicographic incomparability matrix* Q, whose entries are defined as $Q_{ij} = \min(P_{ij}, P_{ji})$ $(i \neq j)$ and $Q_{ii} = 0$. The rationale behind this definition is straightforward: the more one profile dominates the other, the less the two are incomparable (the choice of putting the diagonal elements of Q equal to 0 is to assure the maximum incomparability degree between two profiles to be 0.5: this choice has no essential consequences on the subsequent developments). By the above definition, it follows immediately that Q is symmetric.

3 The scoring functions

As in [7], we assign global dominance (*dom*) and incomparability (*inc*) scores to each profile in the input poset, based on the singular value decompositions of P and Q. However, in order to take into account the distribution of units on π, here we weight P and Q with the relative frequencies over poset profiles. Let s_i be the share of statistical units with profile $\boldsymbol{x}_i$ and let $S = diag(s_1, \ldots, s_m)$. The i-th column of SP comprises the LMRPs of $\boldsymbol{x}_i$, with respect to all of poset profiles, multiplied by the relative frequency corresponding to each of them. In other words, it comprises the probabilities of randomly and independently extracting a statistical unit u and a LLE λ, such that the profile of u is $\boldsymbol{x}_j$ ($j = 1, \ldots, m$) and $\boldsymbol{x}_j$ is not above $\boldsymbol{x}_i$, in λ. Now, let $SP = UD_PV^T$ be the singular value decomposition of matrix SP. From $D_PV^T = U^TSP$, we see that the i-th component dom_i of the first row of D_PV^T is a weighted sum (with non-negative weights, by the Perron-Frobenius Theorem) of such probabilities and can be then interpreted as an overall dominance score of a unit with profile $\boldsymbol{x}_i$, over the other units. As proved in [7][1], the map associating to $\boldsymbol{x}_i$ the dominance score dom_i is *strictly-order preserving* and can serve as a *dominance score function*. Similarly, we take, as global incomparability degrees associated to poset profiles, the components of the first row of D_QB^T, in the singular value decomposition $SQ = AD_QB^T$. In summary, to each profile $\boldsymbol{x}_i$ of the input poset there corresponds a pair (inc_i, dom_i) of incomparability-dominance scores, computed as the i-th components of the first rows of D_QB^T and D_PV^T, respectively. Finally, each statistical unit in the dataset inherits the score pair of its profile and can be mapped into an incomparability-dominance plane, getting both a ranking and a picture of the stretching of the data.

4 Ranking financial literacy in Italy

We have applied the procedure outlined above to data pertaining to financial literacy of Italian adults, collected by Bank of Italy for year 2018 [3]. The survey involved about 2500 respondents, investigating financial knowledge by means of 7 questions, pertaining to different financial notions[2], whose answers are simply classified as *right* or *wrong*. The resulting product order is called 2^7 and has in fact $2^7 = 128$ binary profiles, with $7! = 5040$ lexicographic linear extensions. Dominance and incomparability matrices have been built weighting the LMRPs by the share of respondents in each knowledge profile, also considering the sample weight of each unit. By performing singular value decompositions, each unit has been finally associated the incomparability-dominance score pair. Figure 1 and 2 reproduce the distributions of such pairs for age classes 18-50 and 51+, spatially smoothed to give a more realistic impression of the score patterns. The two plots are scaled to 0-1 on the dominance axis, and the incomparability axis is scaled to 0-0.3 instead due to the small values observed. Units answering correctly all of the questions are scored 1 on the dominance axis, while those providing all wrong answers are scored 0. The value 1 on the incomparability axis corresponds to the theoretical maximum incomparabilty score of a statistical unit, over all possible distributions on the input poset

[1] The pre-multiplication by S does not affect the proof, since all elements of a row are multiplied by the same non-negative quantity.

[2] Questions check knowledge about *inflation*, *simple/compound interest* and *risk management*.

2^7. Consequentially, the small values reported on the incomparability axis mean that the respondents can be reasonably ranked since most of the observed profiles are "almost" comparable. In both age classes the density tends to be concentrated on middle-hìgh dominance scores (more black areas), with evidences of some polarization and the existence of a group of financially illiterate units, separated by the rest of the distribution. Interestingly, older people seem to have more financial competencies than younger subjects; given the increasing relevance of financial products and services for personal lifelong economic sustainability, this appears as a critical feature of the Italian financial literacy distribution. Finally, notice that having similar incomparability degrees need not imply having similar answer profiles, so that care must be taken in interpreting the plot as revealing homogeneous respondent clusters.

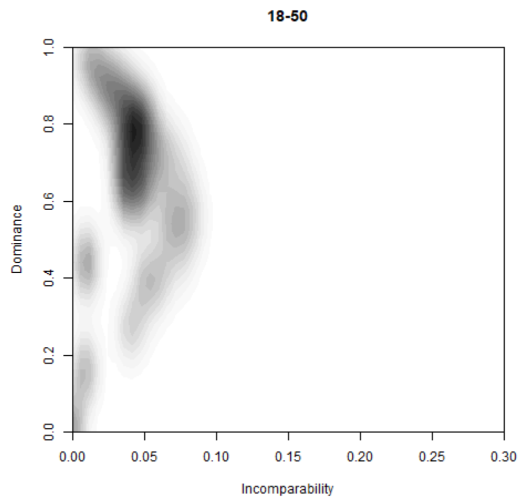

Fig. 1 Smoothed distribution of dominance and incomparability scores related to financial literacy in Italy, for age 18-50. Grey intensity is proportional to frequency density.

5 Conclusions

In this short paper, we have presented a ranking procedure for ordinal multi-indicator systems, which: (i) fully respects the ordinal nature of the data and does not involve any scaling or aggregation of ordinal variables, (ii) is computationally light and can be applied to datasets with a quite large number of variables (possibly, using approximations to the factorial in the computation of LMRPs) and (iii) also provides measures of incomparability among units, so as to keep control on the

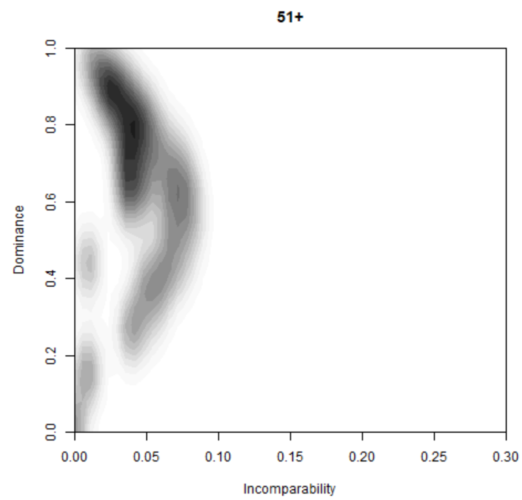

Fig. 2 Smoothed distribution of dominance and incomparability scores related to financial literacy in Italy, for age 51+. Grey intensity is proportional to frequency density.

quality of the ranking process which, as any dimensionality reduction algorithm, unavoidably introduces some distortion into the final data representation. The procedure is freely available in the R package Parsec [5].

References

1. Rainer Brüggemann and Ganapati P Patil. Multicriteria prioritization and partial order in environmental sciences. *Environmental and Ecological Statistics*, 17(4):383–410, 2010.
2. Brian A Davey and Hilary A Priestley. *Introduction to lattices and order*. Cambridge University press, 2002.
3. Antonietta Di Salvatore, Francesco Franceschi, Andrea Neri, and Francesca Zanichelli. Measuring the financial literacy of the adult population: the experience of Banca d'Italia, in "Questioni di Economia e Finanza (Occasional Papers)", n. 435. *Economic Research and International Relations Area*, 2018.
4. Marco Fattore. Partially ordered sets and the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, 128(2):835–858, 2016.
5. Marco Fattore and Alberto Arcagni. Parsec: an R package for poset-based evaluation of multidimensional poverty. In *Multi-indicator systems and modelling in partial order*, pages 317–330. Springer, 2014.
6. Marco Fattore and Alberto Arcagni. A reduced posetic approach to the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, 136(3):1053–1070, 2018.
7. Marco Fattore, Alberto Arcagni, and Filomena Maggino. Optimal scoring of partially ordered data, with an application to the ranking of smart cities. In Giuseppe Arbia, Stefano Peluso, Alessia Pini, and Giulia Rivellini, editors, *Smart Statistics for Smart Applications - Book of Short Papers SIS 2019*, pages 855–860. Pearson, 2019.