Università di Pisa | Sant'Anna Scuola Universitaria Superiore Pisa | Consiglio Nazionale delle Ricerche | Università degli Studi di Cagliari

Book of Short Papers SIS 2021

SIS2021 PISA, 21-25 GIUGNO 2021

SIS Società Italiana di Statistica

Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo

BANCA DI PISA E FORNACETTE CREDITO COOPERATIVO
Banca aderente al Gruppo Bancario Cooperativo Iccrea

TStat
Distribuzione Software | Formazione Professionale
Statistica | Economia | Finanza | Biostatistica | Epidemiologia
Sanità Pubblica | Scienze Sociali
www.tstat.it | www.tstattraining.eu

Copyright © 2021

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

Contents

6 Satellite events 1660

POSetR: a new computationally efficient R package for partially ordered data

POSetR: un nuovo pacchetto R, ad alta afficienza computazionale, per dati parzialmente ordinati

Alberto Arcagni, Alessandro Avellone, Marco Fattore

Abstract In this paper, we introduce `POSetR`, a new `R` package providing highly efficient routines for the treatment of partially ordered data. After motivating the need for a new package on posets, we describe the main functionalities of `POSetR` and give hints on its possible uses.
Abstract *Obiettivo di questo articolo è presentare il nuovo pacchetto* `POSetR`*, per il trattamento di dati parzialmente ordinati, in ambiente* `R`*. Dopo averne motivato la necessità, l'articolo descrive brevemente le principali funzionalità del pacchetto e ne indica i possibili utilizzi.*

Key words: Linear extensions, Mutual ranking probabilities, Partially ordered set, R.

1 Introduction

In this paper, we introduce `POSetR`, a new and efficient `R` [17] package for the analysis of partially ordered data. The package combines high level `R` instructions, with low-level core routines implemented in `C++`, so as to preserve user-friendliness, still assuring for high computational performances. In the following, we (i) briefly discuss the statistical relevance of partially ordered data, (ii) introduce existing `R`

Alberto Arcagni
Dipartimento Metodi e Modelli per l'Economia, il Territorio e la Finanza - Universita di Roma La Sapienza, Via Del Castro Laurenziano 9 - 00161, ROMA. e-mail: alberto.arcagni@uniroma1.it

Alessandro Avellone
Dipartimento di Statistica e Metodi Quantitativi - Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 MILANO. e-mail: alessandro.avellone@unimib.it

Marco Fattore
Dipartimento di Statistica e Metodi Quantitativi - Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 MILANO. e-mail: marco.fattore@unimib.it

resources for posetic analysis, (iii) provide an overview of the new package and (iv) apply it to data of financial knowledge in Italy.

2 Why partially ordered sets?

Many problems in data analysis involve the treatment of multidimensional systems of ordinal indicators., e.g. for the construction of rankings and synthetic indicators, in contexts like the evaluation of multidimensional poverty, quality-of-life or customer satisfaction. Each ordinal indicator provides a *linear order* (possibly with ties) of the statistical units; in general, however, units are ordered differently by different indicators, preventing them to be “globally” ordered. However, they might be *partially* ordered. Indeed, if unit *a* gets “better” scores than unit *b* on all of the indicators, then *a* “dominates” *b*. Thus, some pairs of units can be ordered, some others cannot, producing a partially ordered set, or a *poset* (see [4] and Figure 1). Posets are the natural mathematical structure associated to ordinal multi-indicator systems and, in general, to data described in terms of comparabilities and incomparabilities (e.g. data on preferences); as such, they are also the natural setting to develop a sound “multidimensional ordinal data analysis”, as investigated in recent methodological advances [2, 11, 12, 13, 14, 15].

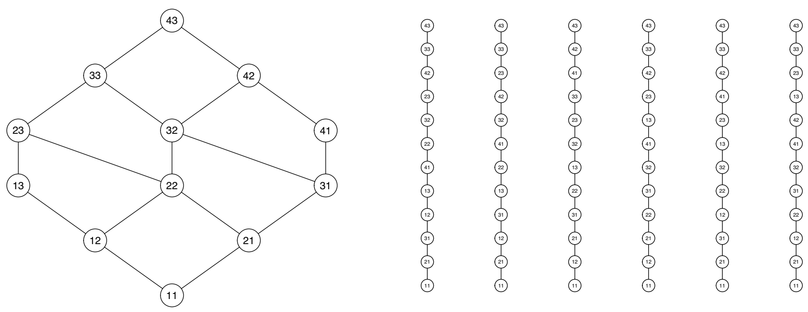

Fig. 1 Hasse diagram of a poset built on two ordinal indicators with 4 and 3 degrees and examples of linear extensions.

3 Computational issues

Posets are combinatoric objects and often generate computationally non-trivial problems. In particular, many posetic data analysis procedures are essentially based on the construction of *linear extensions* (LEs) and on the computation of so-called *mutual ranking probabilities* (MRPs), two related concepts which in general involve heavy computations. The notion of LE of a poset is illustrated in Figure 1. Given a poset π built on a set X, a LE ℓ of π is a linear order of the elements of X such that if $a < b$ in π, then $a < b$ in ℓ; more expressively, ℓ is obtained by resolving all the incomparabilities of π, without affecting the dominances in it. It can be proven that any finite poset can be reconstructed from the set of its LEs [18], which then make it possible to decompose statistical problems on complex partial order structures, into "subproblems" on simple linear orders. For example, ranking extraction, scoring and evaluation over multidimensional ordinal indicator systems [2, 11, 15] are based on the computation of MRPs which in turn involves LE generation, MRPs being the fraction of LEs where an element, say a, dominates another element, say b [14, 15]. Since posets usually encountered in real applications have an extremely high number of LEs, these must be sampled [3]; sampled LEs can nevertheless be of the order of billions, making it crucial to have efficient implementations of sampling algorithms.

4 R resources for posetic analysis

Currently, there is just one `R` package devoted to posets, namely `parsec` [10], developed as a first software implementation of the procedures described in [11, 12, 13, 14, 15]. It implements a quite wide set of tools for basic mathematical analysis of partially ordered sets and for the statistical analysis of partially ordered data, but it is mainly designed to deal with posets built out of multidimensional indicator systems and has its major focus on multi-criteria evaluation. Although it provides quite efficient LE generation algorithms, imported from package `netrankr` [5] (which is not devoted to posets, but implements such algorithms for other purposes), `parsec` is not flexible enough, to effectively adapt to the incoming methodological advances and to the increasing range of statistical applications of posetic tools.

5 Overview of POSetR

`POSetR` is a new posetic package, internally written in C++ and integrated with `R` *via* the `RCpp` package ([7], [8]), developed to provide a well-designed, efficient and flexible "engine" for LE generation. To introduce it, we show some of its main functionalities in action. To create a poset, the set of dominances between objects must be set, as a two-column matrix where elements in the first column are dominated by corresponding elements in the second one, and passed to the poset constructor; function `summary` then provides synthetic infos on the poset:

```
> library(POSetR)
> dom <- matrix(c( "a", "b", "c", "b", "b", "d","e", "c" ),
+      ncol = 2, byrow = TRUE)
> p <- poset(dom, elements = c("a", "b", "c", "d", "e") )
> summary(p)
    5 elements
    8 strict comparabilities
    2 incomparabilities
```

A function `print` is also available, to get the list of dominances of the input poset. The Hasse diagram of the poset is provided by function `plot` (see Figure 2). There are also special functions to generate specific types of posets: `chain` for chains (i.e. linear orders) and `productOrder` for the product order of two posets (tipically used to turn a multi-indicator system into a poset). The key function of the package is `LEapply`. It generates the LEs of the input poset, employing a high performance `C++` code, *at the same time* evaluating an argument function on each extension and taking the average over the set of LEs. The function passed as an argument to `LEapply` can be any user-defined `R` function or a function implemented in the `C++` library of the package (the fundamental function `MutualRankingProbability` is one of them):

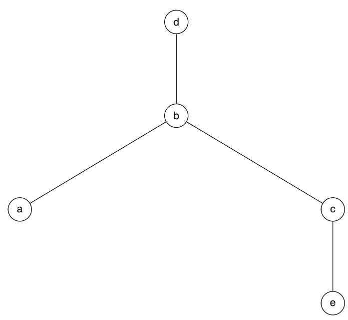

Fig. 2 Hasse diagram of poset `p` produced by the `plot` command.

```
> LEapply(x = p, FUN = "MutualRankingProbability")
    a    b    c    d    e
a 1.00 1.00 0.66 1.00 0.33
b 0.00 1.00 0.00 1.00 0.00
c 0.33 1.00 1.00 1.00 0.00
d 0.00 0.00 0.00 1.00 0.00
e 0.66 1.00 1.00 1.00 1.00
```

`LEapply` makes `POSetR` much more flexible than existing posetic packages, which do not allow the easy implemention of user-defined functions over LEs. This is quite limiting, since many posetic applications to data analysis involve the computation of various statistics over LE, something that in `POSetR` can be efficiently done with a single call to `LEapply`, by properly choosing the argument function (see also the example in Section 6). As mentioned, real posets usually have an extremely high number of LEs, therefore `LEapply` implements both a state-of-the-art algorithm for full LE generation [16] and an MCMC algorithm for LE quasi-uniform sampling (see Bubley and Dyer [3]) and allows users to choose between the two. Interestingly, based on simulations, the computational performances of `POSetR` are of the same order of magnitudeas of those of `netrankr`, if not better.

6 Example: financial knowledge in Italy

We consider the data on financial knowledge provided by Bank of Italy, for year 2020 [1] which comprise the binary scores (0 - No knowledge; 1 - Knowledge) of 2036 individuals, aged 18-79, on 7 financial knowledge dimensions, namely: k_1 *- Purchasing power;* k_2 *- Interest paid on a mortgage;* k_3 *- Simple interest calculation;* k_4 *- Compound interest calculation,* k_5 *- Risk and reward;* k_6 *- Inflation definition*; k_7 *- Diversification*. The $2^7 = 128$ possible binary score patterns (knowledge *profiles*) are partially ordered *componentwise* and to each of them the corresponding relative frequency is associated. To summarize the data, we want to identify the *median* knowledge pattern of the population. The partial order structure of the data, however, makes the notion of the median a fuzzy one [9]; the "degree of membership to the median" of a profile is computed as the fraction of LEs of the financial knowledge poset in which it represents the median (indeed, linear extensions are completely ordered and on the median can be defined in the classical way). Knowledge profiles are first enumerated from 1 to 128, then `LEapply(p, FUN = isMedian, generator = "BubleyDyer", bubleydyer.nit = n^3)` is called, where `isMedian` is a function returning, for each LE, a binary vector of length 128, whose *i*-component is 1 if and only if the *i*-th profile is the median in that LE (given the number of profiles, the Bubley-Dyer MCMC algorithm has been used). More than 40 profiles happen to represent the median pattern is some LEs, but most of them have very small membership degrees (*md*, for short). The three profiles with the highest degrees are 0110110 (*md*: 0.23), 1100110 (*md*: 0.11) and 0100111 (*md*: 0.11). Such profiles can be considered as representative of the main features of financial knowledge in Italy. Interestingly they share the lack of knowledge about *compound interest* (4th component) and *diversification* (7th component), two dimensions of key importance for financial awareness.

7 Conclusion

We have presented the new R package POSetR and its basic functionalities, particularly the highly efficient C++ implementation of the "engine", for generating linear extensions and flexibly computing user-defined functions on them. Future software implementations will move along two lines: (i) improving POSetR adding new functionalities, mainly oriented to the mathematical treatment of partial order structures and (ii) developing and integrating procedures for the statistical analysis of partially ordered data, to provide an ecosystem for applied statisticians.

References

1. D'Alessio G., De Bonis R., Neri A. Rampazzi C. (2020) *Financial literacy in Italy: the results of the Bank of Italy's 2020 survey*, Bank of Itali . Occasional Papers 588.
2. Bruggemann R., Patil G. P. (2011). *Ranking and Prioritization for Multi-indicator Systems*, Springer.
3. Bubley R., Dyer M. (1999). *Faster random generation of linear extensions*, Discrete Math., 201, 81-88.
4. Davey B. A., Priestley B. H. (2002). *Introduction to Lattices and Order*, CUP.
5. David Schoch (2017). *netrankr: An R package to analyze partial rankings in networks*.
6. De Loof K. (2010). *Efficient computation of rank probabilities in posets*, Ph.D dissertation, available on https://biblio.ugent.be/publication/874495.
7. Dirk Eddelbuettel and Romain Francois (2011). Rcpp: *Seamless R and C++ Integration*, Journal of Statistical Software, 40(8), 1-18. URL https://www.jstatsoft.org/v40/i08/.
8. Dirk Eddelbuettel and James Joseph Balamuta (2018). *Extending R with C++: A Brief Introduction to Rcpp*, The American Statistician. 72(1). URL https://doi.org/10.1080/00031305.2017.1375990.
9. Fattore M. (2008). *Hasse diagrams, poset theory and fuzzy poverty measures*, Rivista internazionale di scienze sociali 1, 63-75..
10. Fattore M., Arcagni A. (2014). *PARSEC: an R package for poset-based evaluation of multidimensional poverty*, in: Bruggemann R., Carlsen L., Wittmann J. (Eds.) Multi-Indicator Systems and Modelling in Partial Order, Springer, Berlin.
11. Fattore M. (2016). *Partially ordered sets and the measurement of multidimensional ordinal deprivation*, Social Indicators Research. 128(2) pp. 835-858 2, DOI: 10.1007/s11205-015-1059-6.
12. Fattore M., Arcagni A. (2018). *F-FOD: Fuzzy First Order Dominance analysis and populations ranking over ordinal multi-indicator systems*, Social Indicators Research DOI: 10.1007/s11205-018-2049-2.
13. Fattore M., Arcagni A. (2018). *A reduced posetic approach to the measurement of multidimensional ordinal deprivation*, Social Indicators Research, 136(3), 1053-1070.
14. Fattore M., Arcagni A., Maggino F. (2019). *Optimal scoring of partially ordered data, with an application to the ranking of smart cities*, Smart Statistics for Smart Applications Book of Short Papers SIS 2019 - ISBN 9788891915108.
15. Fattore M., Arcagni A. (2020). *Ranking extraction in ordinal multi-indicator systems*, Book of Short Papers - SIS 2020 - ISBN 9788891910776 - Pearson.
16. Habib, M., Medina, R., Nourine, L., and Steiner, G. (2001). *Efficient algorithms on distributive lattices*, Discrete Applied Mathematics, 110(2-3), 169-187.
17. R Core Team (2020). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
18. Schröder. (2002). *Ordered set. An introduction*, Birkäuser.