Relational data analysis for weakly structured information: Utilizing linear and binary programming for computing supremum statistics on closure systems

Georg Schollmeyer Christoph Jansen Thomas Augustin

05.07.2017

Department of Statistics, Ludwig-Maximilians-Universität, Munich

Relational data analysis

Numerical data analysis

A given set $\mathcal{O} = \{o_1, \ldots, o_m\}$ of objects (data points, statistical units) is analyzed by analyzing numerical assignments $u(o_1), \ldots, u(o_m)$. (e.g., person o_i has an income of 1200 Euro).

Relational data analysis

A given set $\mathcal{O} = \{o_1, \ldots, o_m\}$ of objects (data points, statistical units) is analyzed by either

- analyzing empirical relations R between the objects (e.g., does person o_i have a smaller income than person o_j, $\rightsquigarrow$ order theory) or
- analyzing empirical relations I between the objects and certain attributes $\mathcal{A} = \{a_1, \ldots, a_m\}$ (does object o_i have attribute a_j, e.g., is person o_i male?, $\rightsquigarrow$ e.g., Formal concept analysis (FCA, [Ganter and Wille, 2012])).

Imprecise/Relational data

- Epistemic case: precise data point x that can only be observed in interval form $[x_l, x_u]$ and one only knows $x \in [x_l, x_u]$.
  Example: imprecisely observed data about income in social surveys (to reduce item non-response rate).

ALLBUS 2014: Variable Report

GESIS Studien-Nr. 5240 (v2.1.0), http://dx.doi.org/10.4232/1.12288

ZA5240, V418: (N=328) (gewichtet nach V870)

V418

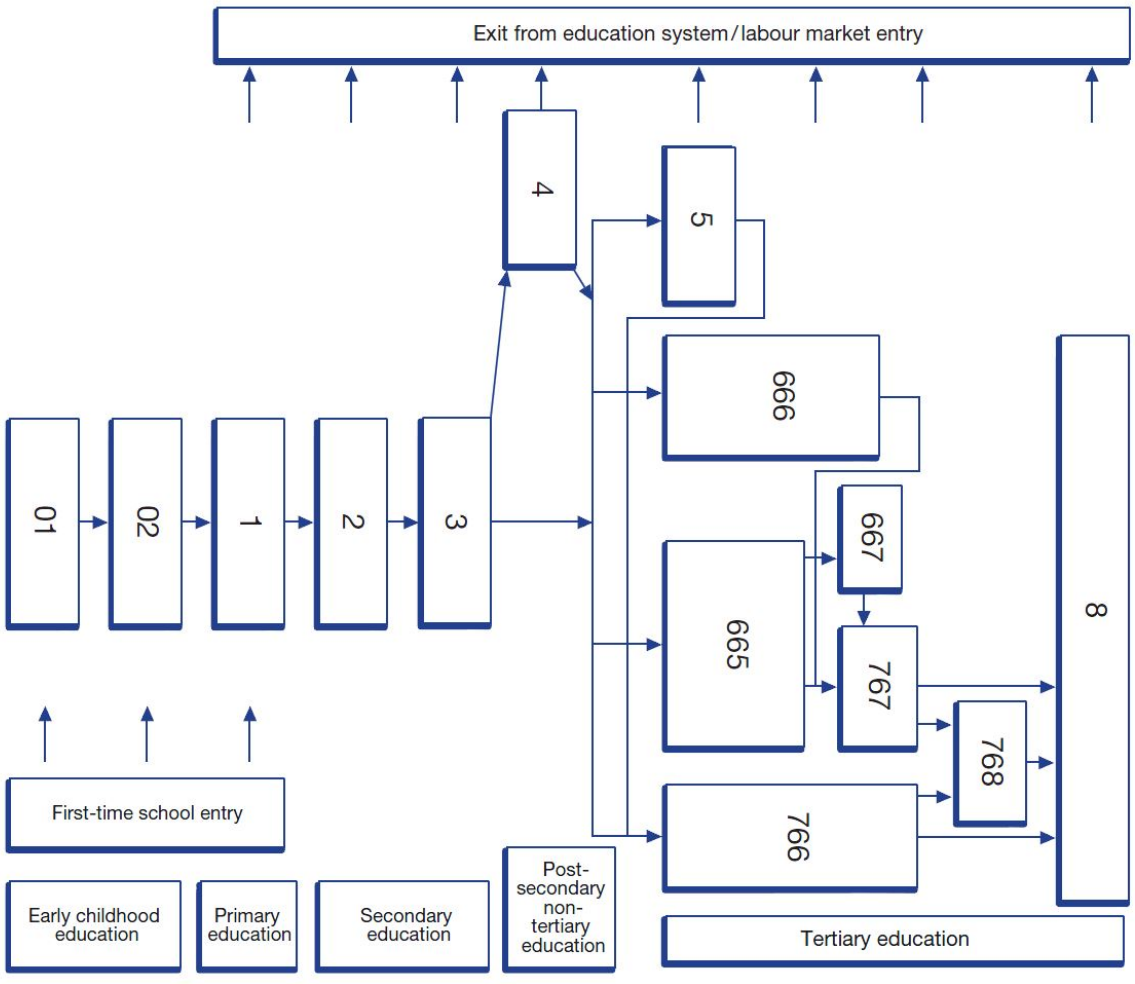
Wert	Ausprägung	Missing	Anzahl	Prozent	Gült.Prozent
0	KEIN EINKOMMEN	M	208	6,0	
1	UNTER 200 EURO		4	0,1	1,2
2	200 - 299 EURO		5	0,1	1,5
3	300 - 399 EURO		5	0,1	1,5
4	400 - 499 EURO		5	0,1	1,5
5	500 - 624 EURO		5	0,1	1,5
6	625 - 749 EURO		14	0,4	4,3
7	750 - 874 EURO		16	0,5	4,9
8	875 - 999 EURO		18	0,5	5,5
9	1000 - 1124 EURO		20	0,6	6,1
10	1125 - 1249 EURO		9	0,3	2,7
11	1250 - 1374 EURO		15	0,4	4,6
12	1375 - 1499 EURO		19	0,5	5,8

$o_i < o_j \iff o_i$ is in a lower income-category than o_j.

Imprecise/Relational data

- Ontic case: Data point x is observed precisely, but has no numerical, but only an ordinal/relational character.
  Example: Formal education in poverty analysis.

Example: Different educational paths in International Standard Classification of Educationin [UNESCO Institute for Statistics (UIS), 2012]



ISCED-Programmes (ISCED-P)	
0	Early childhood education
1	Primary education
2	Lower secondary education
3	Upper secondary education
4	Post-secondary non-tertiary education
5	Short-cycle tertiary education
6	Bachelor's or equivalent level
7	Master's or equivalent level
8	Doctoral or equivalent level
9	Not elsewhere classified

$o_i < o_j \iff o_i$ followed the same educational path like o_j, but stopped earlier.

Relational data analysis

- Relational data analysis applicable for both cases,
  but appears more naturally for ontic case.
- But: for the example of detecting stochastic dominance
  for interval-valued data (later):
  Ontic type analysis technically easier and will lead to
  exactly the same results like an epistemic type analysis.

Basic situation

- Basic set V.
- Closure system $\mathcal{S} \subseteq 2^V$ (i.e., a family $\mathcal{S}$ of subsets of V that contains V and that is closed under arbitrary intersections).
- Here: discrete case: V finite.
- Aim: compute supremum type statistic

$$D^+ := \sup_{S \in \mathcal{S}} f(S)$$

where

$$f(S) = \langle w, \mathbb{1}_S \rangle$$

is a linear function in the indicator function of S.

Supremum statistics on closure systems: Examples of application

- Multivariate generalizations of Kolmogorov-Smirnov test:
  - Spatial statistics (closure system of all convex areas)
  - Item response theory (IRT): multivariate item impact and Differential item functioning analysis (DIF) in dichotomous IRT- datasets (closure system of all principal filters)
  - Formal concept analysis / Subgroup discovery analysis (closure system of all formal concept extents/ closure system of all subgroups)
- Multivariate generalizations of first order stochastic dominance (closure system of all upsets of a partially ordered set)

Example 1: Two sample Kolmogorov-Smirnov test / test of stochastic dominance

- $V = \mathbb{R}$
- Two samples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$ with associated empirical measures $\hat{P}_x$ and $\hat{P}_y$
- Closure system $\mathcal{S} = \{ [c, \infty) \mid c \in \mathbb{R} \}$ of all upsets of $(\mathbb{R}, \leq)$
- statistic $D^+ := \sup_{[c,\infty) \in \mathcal{S}} \hat{P}_x([c, \infty)) - \hat{P}_y([c, \infty))$,
- or statistic $D^- := \inf_{[c,\infty) \in \mathcal{S}} \hat{P}_y([c, \infty)) - \hat{P}_x([c, \infty))$
- or statistic $D := \max\{D^+, -D^-\}$
- It is enough to look at the finite closure system $\mathcal{S}_{|x,y} := \{ S \cap \{x_1, \ldots, x_n, y_1, \ldots, y_m\} \mid S \in \mathcal{S}\}$

Example 2: Kolmogorov-Smirnov test and Stochastic dominance in higher dimensions

- $\mathbb{V} = (V, \leq)$ **partially** ordered set (poset)
- Two samples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$ with associated empirical measures $\hat{P}_x$ and $\hat{P}_y$
- Closure system $\mathcal{S} = \{\uparrow x := \{y \mid y \geq x\} \mid x \in V\}$ of all principal filters of $\mathbb{V}$ for Kolmogorov-Smirnov test.
- Closure system $\mathcal{S} = \{\ A \subseteq V \mid a \in A \ \&\ b \geq a \Longrightarrow b \in A\ \}$ of all upsets of $(V, \leq)$ for Stochastic dominance
- statistic $D^+ := \sup_{S \in \mathcal{S}} \hat{P}_x(S) - \hat{P}_y(S)$ (and D^- and D)
- Closure system $\mathcal{S}_{|x,y}$ typically very large.

Example 3: Item impact and Differential item functioning (DIF) in dichotomous item response data

- Given set $\mathcal{A} = \{a_1, \ldots, a_n\}$ of items (e.g., questions) of an IRT test battery that was solved by a set $\mathcal{O} = \{o_1, \ldots, o_m\}$ of persons.
- Given incidence relation $I \subseteq \mathcal{O} \times \mathcal{A}$ with $(o_i, a_j) \in I :\Longleftrightarrow$ person o_i answered question a_j correctly.
- For subset $A \subseteq \mathcal{A}$ define the set $Y_A := \{o \in \mathcal{O} \mid \forall a \in A : (o, a) \in I\}$ of all persons that answered at least all questions in A correctly.
- Then, the family $\mathcal{S} := \{Y_A \mid A \subseteq \mathcal{A}\}$ of all such sets of persons is a closure system. (In FCA it is called the closure system of all formal concept extents)

Example 3: Item impact and Differential item functioning (DIF) in dichotomous item response data

- Given a target attribute (e.g., gender) one can look at every such set and compute that set, for which the target variable has the most unusual statistical (distributional) characteristics, (e.g., the difference in proportions of male and female persons is very large)
- This is exactly the problem of computing a supremum of a linear function over a closure system
- Closely related to problem statement of subgroup discovery analysis (e.g., [Klösgen, 1996, Wrobel, 1997, Lavrač et al., 2004, Atzmueller, 2015], there: more than one subgroup, different, generally non-linear objective functions)

Solving the optimization problem $D^+ = \sup_{S \in \mathcal{S}} \langle w, \mathbb{1}_S \rangle$

- Typically, the closure system is very large $\rightsquigarrow$ explicit computation of $\langle w, \mathbb{1}_S \rangle$ for all sets of the closure system not feasible
- One can formulate the optimization problem as a binary linear program.
- The demand $S \in \mathcal{S}$ can be modeled by using contextual logic from FCA. Example: For the closure system of all upsets: If $a \leq b$, then every upset that contains a necessarily must also contain b.
- The closure system of all upsets is exactly described by all such formal implications
- All these formal implications can be implemented via inequality constraints of the form $\mathbb{1}_S(a) \leq \mathbb{1}_S(b)$.

Solving the optimization problem $D^+ = \sup_{S \in \mathcal{S}} \langle w, \mathbb{1}_S \rangle$

- Case of stochastic dominance: demand of decision variables to be binary can be dropped $\rightsquigarrow$ Classical linear program that is easy enough to solve.
- Case of item impact/DIF: One can model the formal implications between questions and persons (e.g.: if o_i is in Y_A and if o_i did not solve a_j, then a_j cannot be in A).
- This leads to binary program with $m + n$ binary decision variables and $\mathcal{O}(m + n)$ constraints, where for n decision variables the demand of being binary can be dropped.
- Obtained program can also be used in the context of subgroup discovery.

Statistical inference for $D^+ = \sup_{S \in \mathcal{S}} \langle w, \mathbb{1}_S \rangle$

- Teststatistic D^+ not distribution-free
- permutation test
- Vapnik-Chervonenkis theory ([Vapnik and Chervonenkis, 1968, 1971]) for large deviation bounds:
  Neat relations between V.C.-dimension and order theoretic/lattice-theoretic notions (e.g., for upsets: V.C.-dimension = width, for distributive lattices: V.C.dimension of all principal filters = order dimension) .
- Vapnik-Chervonenkis theory also supplies possibility of regularization.

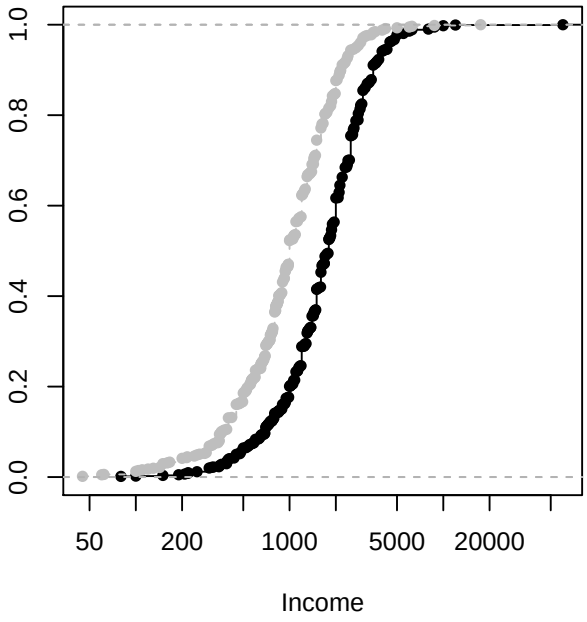
Summary

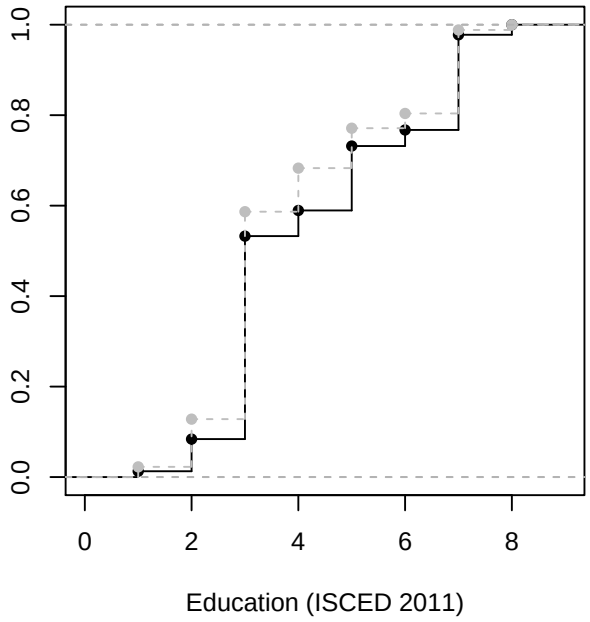
- We introduced linear/binary programs for computing supremum statistics on closure systems.
- Areas of application include:
  - Spatial statistics
  - Multivariate stochastic dominance (e.g., in poverty analysis)
  - Item impact/DIF in item response theory
  - generally: Subgroup discovery / Formal concept analysis
- We can solve the inference problem with permutation tests
- Additional statistical analysis with V.C.-theory (including computing and trimming V.C.-dimension $\rightsquigarrow$ regularization)

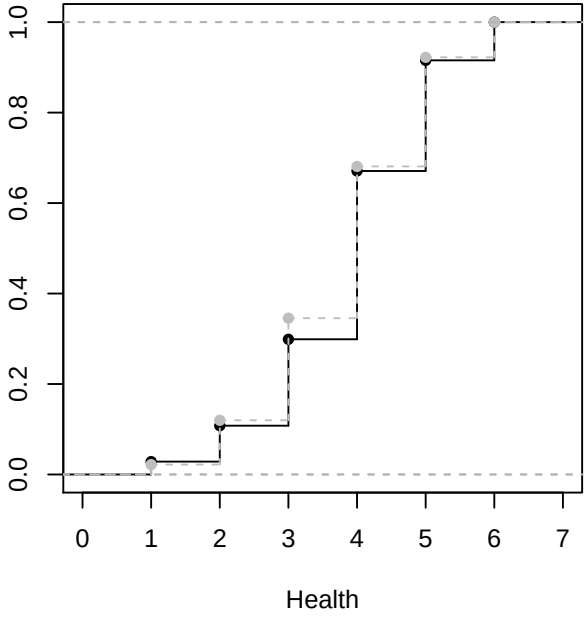
Application example

- Subsample of Allbus 2014 (706 female and 809 male respondents).
- Dimensions:
  - *Income*.
  - *Health* (self-reported, ranging from 1 (bad) to 6 (excellent)).
  - *Education* (ISCED 2011: ranging from 0 (less than primary education) to 8 (doctoral or equivalent level)).

Marginal analysis

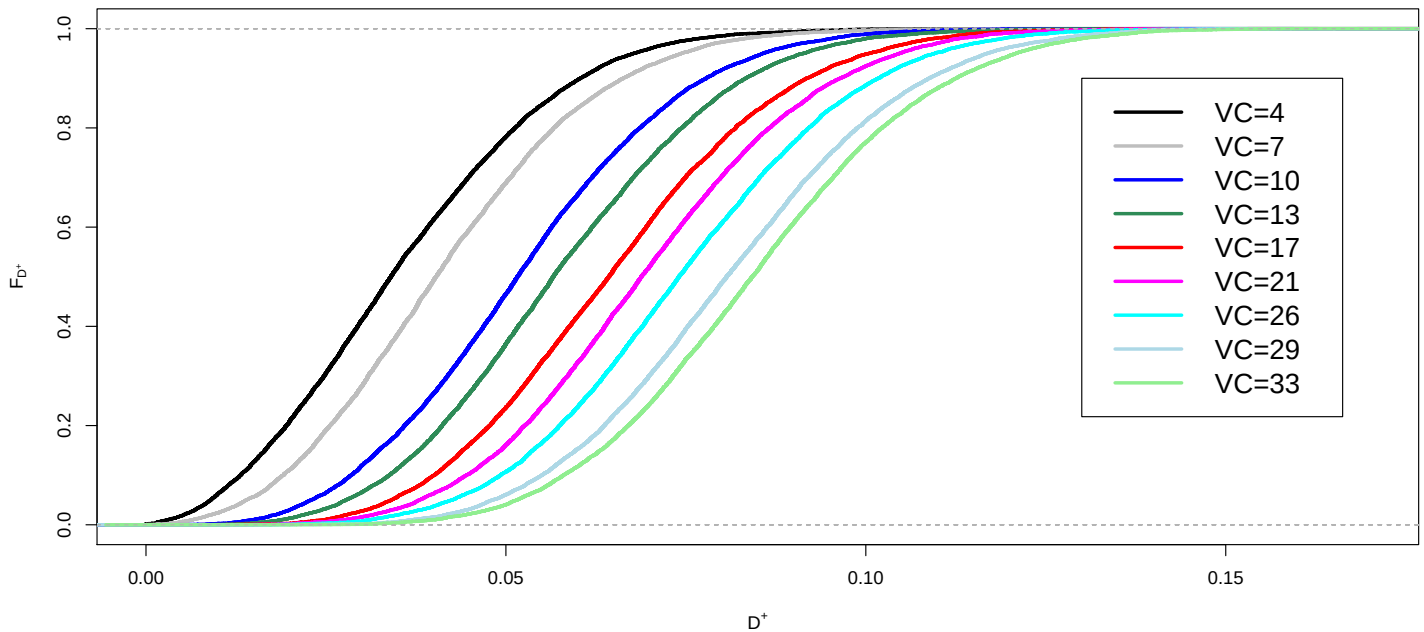






Joint analysis

- V.C.-dim $= 33$ (number of upsets $\in [10^{10}, 10^{60}]$, dual simplex algorithm took less than a second).
- $D^+ \approx 36.5\%$.
- $D^- \approx -1.2\%$.
- Female subgroup (Y) almost stochastically smaller than male subgroup (X).
- Value of
  D^+ significantly positive, D^- not significantly different from zero.



References

M. Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1): 35–49, 2015. doi: 10.1002/widm.1144. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1144`.

B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 2012.

W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*, pages 249–271. American Association for Artificial Intelligence, 1996.

N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5(Feb):153–188, 2004.

UNESCO Institute for Statistics (UIS). *International standard classification of education: ISCED 2011*. UIS, Montreal, Quebec, 2012.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Proceedings of the USSR academy of sciences*, 181(4): 781–783, 1968.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

S. Wrobel. An algorithm for multi-relational discovery of subgroups. In J. Komorowski and J. Zytkow, editors, *Principles of Data Mining and Knowledge Discovery*, pages 78–87, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. ISBN 978-3-540-69236-2.