LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Georg Schollmeyer, Christoph Jansen, Thomas Augustin

Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems

Technical Report Number 209, 2017
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems

Georg Schollmeyer Christoph Jansen Thomas Augustin

Abstract

In this paper we develop a linear programming method for detecting stochastic dominance for random variables with values in a partially ordered set (poset) based on the upset-characterization of stochastic dominance. The proposed detection-procedure is based on a descriptively interpretable statistic, namely the maximal probability-difference of an upset. We show how our method is related to the general task of maximizing a linear function on a closure system. Since closure systems are describable via its valid formal implications, we can use here ingredients of formal concept analysis. We also address the question of inference via resampling and via conservative bounds given by the application of Vapnik-Chervonenkis theory, which also allows for an adequate pruning of the envisaged closure system that allows for the regularization of the test statistic (by paying a price of less conceptual rigor). We illustrate the developed methods by applying them to a variety of data examples, concretely to multivariate inequality analysis, item impact and differential item functioning in item response theory and to the analysis of distributional differences in spatial statistics. The power of regularization is illustrated with a data example in the context of cognitive diagnosis models.

Keywords: stochastic dominance, multivariate stochastic order, linear programming, closure system, formal concept analysis, formal implication, Vapnik-Chervonenkis theory, regularization.

1 Introduction

Stochastic (first order) dominance plays an important role in a huge variety of disciplines like for example welfare economics (cf., e.g., [Arndt et al., 2012, 2015]), decision theory (cf., e.g., [Levy, 2015]), portfolio analysis (cf., e.g., [Kuosmanen, 2004]), nonparametric item response theory (IRT, cf., e.g., [Scheiblechner, 2007]), medicine (cf., e.g., [Leshno and Levy, 2004]), toxicology (cf., e.g., [Davidov and Peddada, 2013]) or psychology (cf., e.g., [Levy and Levy, 2002]) to cite just a few. Most treatments of stochastic dominance are devoted to the univariate case with emphasis also on higher order stochastic dominance

or to the classical multivariate case where one has $\mathbb{R}^d$-valued random variables with the natural ordering $\leq = \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mid \forall i \in \{1, \ldots, d\} : x_i \leq y_i\}$. In this paper we treat the general case of random variables that have values in a partially ordered set[1] (poset) $\mathbb{V} = (V, \leq)$.

Detecting stochastic dominance in this general case is especially interesting in the context of multivariate inequality or poverty analysis (cf., [Alkire et al., 2015]) in the situation where one has more dimensions of inequality that are additionally possibly only of a partial ordinal scale of measurement. One thinkable dimension with an only partially ordered scale of measurement is the dimension education, because different highest educational achievements may be incomparable due to different specifics of different courses of education. In this paper, the example of multivariate inequality analysis will serve as a prototypic example of multivariate stochastic dominance analysis.

In contrast to the simple univariate case, for random variables with values in a partially ordered set the notion of stochastic dominance cannot be simply described with the distribution function, anymore[2]. For two random variables $X : \Omega \longrightarrow V$ and $Y : \Omega \longrightarrow V$ with values in a partially ordered set $(V, \leq)$, one says that X is (weakly) stochastically smaller than Y, denoted by $X \leq_{SD} Y$, if there exist random variables X' and Y' on a further probability space $(\Omega', \mathcal{F}', P')$ with $X \stackrel{d}{=} X', \quad Y \stackrel{d}{=} Y'$ and $P'(X' \leq Y') = 1$. The property of stochastic dominance can be characterized by three essentially equivalent, more constructive statements: The random variable X is stochastically smaller than the random variables Y if one of the three following conditions is satisfied[3]:

i) $P(X \in A) \leq P(Y \in A)$ for every (measurable) upset $A \subseteq V$

ii) $\mathbb{E}(u \circ X) \leq \mathbb{E}(u \circ Y)$ for every bounded non-decreasing Borel-measurable[4] function $u : V \longrightarrow \mathbb{R}$

iii) It is possible to obtain the density[5] f_X from the density f_Y by transporting probability mass from values v to smaller values $v' \leq v$.

In this paper we will deal with the problem of detecting stochastic dominance between two random variables X and Y for which one has observed an i.i.d. sample $(x_1, \ldots, x_{n_x})$

[1] This includes especially the multivariate case of $\mathbb{R}^d$ where the natural order $x \leq y \iff \forall i \in \{1, \ldots, d\} : x_i \leq y_i$ is used. Note also that every finite poset $(V, \leq)$ can mathematically be represented as a multivariate case where the dimension equals the order dimension of $(V, \leq)$, cf. [Dushnik and Miller, 1941, Trotter, 2001].

[2] If one would rely on the distribution function in the multivariate case, then one would get another order, the so-called lower orthant or upper orthant order, cf., e.g., [Müller and Stoyan, 2002].

[3] The equivalence between (ii) and (i) was shown by Lehmann [1955] and independently proved by Levhari et al. [1975]. The equivalence between (iii) and (i) is a consequence of Strassen's Theorem ([Strassen, 1965]), see Kamae et al. [1977].

[4] Here, we have to assume that $(V, \leq)$ can be equipped with an appropriate topology that makes it a partially ordered polish space.

[5] This statement is of course only equivalent if the densities f_X and f_Y actually exist.

of the unknown random variable X and an i.i.d. sample $(y_1, \ldots, y_{n_y})$ of the unknown random variable Y. Actually, one would be interested in detecting $X \leq_{SD} Y$, but one does not exactly know the true law of X and Y. So, here we will deal with detecting empirical stochastic dominance between X and Y, denoted by $X \leq_{\hat{SD}} Y$, where the true laws of X and Y are replaced by the corresponding empirical laws. The problem of statistical inference that is concerned with the question of how stochastic dominance w.r.t. the empirical laws can be translated to stochastic dominance w.r.t. the true laws will also be discussed in this paper. The typical situation in this paper will be the analysis of differences between two subpopulations of some population. The typical subpopulations analyzed in this paper will be subpopulations of male and female persons. Here, we think of the random variable X as the outcome of a random sample from the subpopulation of the male, and Y as a random sample from the subpopulation of the female persons. Note that in the formal definition of stochastic dominance one compares random variables on the same probability space $(\Omega, \mathcal{F}, P)$. In our case of comparing subpopulations we can ensure that X and Y are random variables on the same underlying probability space by thinking of jointly sampling from the male and the female subpopulation. Note that the notion of stochastic dominance does not rely on the possible dependencies between X and Y, because all terms involved in the characterizing properties $i) - iii)$ of stochastic dominance only rely on the marginal distribution of X and Y. Note further that the definition of stochastic dominance could thus be simply extended to random variables living on different probability spaces. Thus, also for the replacement of the true laws by empirical laws, different sample sizes for the male and the female samples would not introduce any problem, here.

For detecting stochastic dominance in the above sense, we will make substantial use of the upset-characterization $i)$. The characterization via a mass transfer can also be used to check for stochastic dominance, see, e.g., Mosler and Scarsini [1991] (for empirical applications see, e.g., [Arndt et al., 2012, 2013]), while an alternative approach would be to make use of a network flow formulation of the problem, as outlined in Preston [1974] or Hansel and Troallic [1978] and then check for dominance via computation of the maximum flow. The main reason for putting emphasis on the upset approach is that with this approach we could not only check for stochastic dominance, but we will also get additionally some well-interpretable statistic for free, upon which we can also base an attempt to do inference. Beyond this, the family of all upsets of a given poset is a well understood closure system[6] and a natural question is then, how the linear programming approach outlined here, can be generalized to the case of arbitrary closure systems.

The paper is structured as follows: In Section 2 we briefly introduce basic mathematical concepts of partially ordered sets, complete lattices and formal concept analysis needed in the paper. Section 3 develops and analyses a linear program for detecting first order sto-

[6] A closure system $\mathcal{S}$ is a family of subsets of a space Ω that contains Ω and is closed under arbitrary intersections.

chastic dominance for random variables with values in a poset. In Section 4 we generalize the linear programming approach to optimization on closure systems. Statistical inference for the developed methods, especially the application of Vapnik-Chervonenkis theory, possible regularization and characterizations of the Vapnik-Chervonenkis dimension of selected closure systems (as well as concretely computing the Vapnik-Chervonenkis dimension) are treated in Section 5. Examples of application, ranging from inequality analysis based on stochastic dominance to a geometrical generalization of the Kolmogorov-Smirnov test for spatial statistics are given in Section 6, while Section 7 concludes.

2 Mathematical preliminaries

In this section, we very briefly introduce elementary basics of partially ordered sets and of formal concept analysis. A far more detailed introduction to partially ordered sets can be found in Davey and Priestley [2002], which also gives a short introduction to formal concept analysis. An introduction into formal concept analysis can be found in Ganter and Wille [2012]. The concepts of formal concept analysis are actually only needed for the optimization problems on general closure systems indicated in Section 4, the reader only interested in the problem of detecting first order stochastic dominance can skip Section 2.2.

2.1 Ordered sets and lattices

Definition 1 (posets and lattices)**.** *A* **partially ordered set (poset)** $\mathbb{V} = (V, \leq)$ *is a pair of a set* V *and a binary relation* $\leq$ *on* V *that is reflexive transitive and antisymmetric. A poset* $(V, \leq)$ *is called* **linearly ordered***, if every two elements* x, y *of* V *are comparable (meaning that* $x \leq y$ *or* $y \leq x$*). For two different elements* x, y *of a poset* $\mathbb{V}$ *we say that* y *is an upper neighbor of* x *(or that* x *is a lower neighbor of* y*), and denote this by* $x \lessdot y$*, if* $x \leq y$ *and if there is no further element* $z \in V$ *(different from* x *and* y*) with* $x \leq z \leq y$*.*

A **lattice** $\mathbb{L} = (L, \leq)$ *is a poset such that every set* $\{x, y\}$ *of two elements* $x, y \in L$ *has a least upper bound and a greatest lower bound. A lattice is called* **complete***, if every arbitrary set* M *has a least upper bound and a greatest lower bound. The least upper bound of a set* M *is called* **supremum** *or* **join** *of* M *and it is denoted with* $\bigvee M$*. The greatest lower bound of a set* M *is called* **infimum** *or* **meet** *of* M *and it is denoted with* $\bigwedge M$*. An element* x *of a complete lattice* $(L, \leq)$ *is called* **join-irreducible** *if for arbitrary subsets* $B \subseteq L$ *from* $x = \bigvee B$ *it follows* $x = b$ *for some* $b \in B$*. The set of all join-irreducible elements of a poset* $\mathbb{V}$ *is denoted with* $\mathcal{J}(\mathbb{V})$*.*

Definition 2 (upset and downset, principal ideal and principal filter)**.** *Let* $(V, \leq)$ *be a poset. A set* $V \subseteq M$ *is called an* ***upset*** *(or filter) if we have* $\forall x, y \in V : x \leq y \ \& \ x \in M \implies y \in M$*. A subset* $M \subseteq V$ *is called* ***downset*** *(or ideal) if* $\forall x, y \in V : x \leq y \ \& \ y \in M \implies x \in M$*. The set of all upsets of a poset* $(V, \leq)$ *is denoted with* $\mathcal{U}((V, \leq))$ *and the set of all downsets is denoted with* $\mathcal{D}((V, \leq))$*. An upset of the form* $\uparrow x := \{y \in V \mid y \geq x\}$

with $x \in V$ is called a **principal filter**. *A downset of the form $\downarrow x := \{y \in V \mid y \leq x\}$ with $x \in V$ is called a* **principal ideal.**

Remark 1. *The complement of an upset is a downset and the complement of a downset is an upset.*

Definition 3 (chain, antichain and width)**.** *Let $(V, \leq)$ be a poset. A set $M \subseteq V$ is called a* ***chain*** *if every two arbitrary elements x and y of M are comparable (meaning that $x \leq y$ or $y \leq x$). A subset M of a poset $(V, \leq)$ is called an* ***antichain*** *if every two arbitrary different elements x and y of M are incomparable (meaning that neither $x \leq y$ nor $y \leq x$). The* **width** *of a finite poset $(V, \leq)$ is the maximal cardinality of an antichain of $(V, \leq)$.*

Remark 2. *For every upset M the set $\min M$ of all minimal elements of M is an antichain. Furthermore, every finite upset M can be characterized by its minimal elements as $M = \uparrow \min M := \{x \in V \mid \exists y \in \min M : y \leq x\}$. Analogous statements are valid for downsets.*

Definition 4 (order dimension)**.** *The* **order dimension** *of a poset $(V, \leq)$ is the smallest number k such that there exist k linearly ordered sets $(V, L_1), \ldots, (V, L_k)$ that represent $(V, \leq)$ via $\leq = \bigcap_{i=1}^{k} L_i$.*

2.2 Formal concept analysis

Formal concept analysis (FCA) is an applied mathematical theory rooted in an attempt to mathematically formalize the notion of a *concept*. In its origins initially motivated by some philosophical attempt to restructure lattice theory (cf., [Wille, 1982]) it nowadays also has very broad applications in computer science, for example in data mining, text mining, machine learning or knowledge management, to name just a few.

Concretely, in formal concept analysis one starts with a so-called **formal context** $\mathbb{K} = (G, M, I)$ where G is a set of objects, M is a set of attributes and $I \subseteq G \times M$ is a binary relation between the objects and the attributes with the interpretation $(g, m) \in I$ iff object g has attribute m. If $(g, m) \in I$ we also use infix notation and write gIm. In the context of statistical data analysis, the objects are often the data points, for example the persons that participated in a survey.The attributes are the observed values of the interesting variables, for example the answer *yes* or *no* to the posed questions and gIm means that person g answered the question m with *yes*. (If the answers to the questions in a survey are not binary, then one can transform them into binary attributes with the method of conceptual scaling, see below.) A **formal concept** of the context $\mathbb{K}$ is a pair (A, B) of a set $A \subseteq G$ of objects, called **extent**, and a set $B \subseteq M$ of attributes, called **intent**, with the following properties:

1. Every object $g \in A$ has every attribute $m \in B$ (i.e.: $\forall g \in A \forall m \in B : gIm$).
2. There is no further object $g \in G \backslash A$ that has also all attributes of B (i.e.: $\forall g \in G : (\forall m \in B : gIm) \Longrightarrow g \in A$).

3. There is no further attribute $m \in M \backslash A$ that is also shared by all objects $g \in A$ (i.e. $\forall m \in M : (\forall g \in A : gIm) \Longrightarrow m \in B$).

Conceptually, the concept extent describes, which objects belong to the formal concept and the intent describes, which attributes characterize the concept. The property of being a formal concept can be characterized with the following operators

$$\begin{aligned} \Phi : 2^M \longrightarrow 2^G : &\quad B \mapsto \{g \in G \mid \forall m \in B : gIm\} \\ \Psi : 2^G \longrightarrow 2^M : &\quad A \mapsto \{m \in M \mid \forall g \in A : gIm\} \end{aligned}$$

as

$$(A, B) \text{ is a formal concept} \iff \Psi(A) = B \ \& \ \Phi(B) = A.$$

This can be verbalized as: "The pair (A, B) is a formal concept iff B is exactly the set of all common attributes of the objects of A and A is exactly the set of all objects having all attributes of B." In the sequel, we will abbreviate both Ψ and Φ with $'$. (Which of the two operators is meant will be always clear from the context.) Furthermore, for singleton sets $\{g\} \subseteq G$ or $\{m\} \subseteq M$ we abbreviate $\{g\}'$ by g' and $\{m\}'$ by m'.

On the set of all formal concepts we can define a subconcept relation as

$$(A, B) \leq (C, D) \iff A \subseteq C \ \& \ B \supseteq D.$$

(Actually, for formal concepts the equivalence $A \subseteq C \iff B \supseteq D$ holds.) If the concept (A, B) is a subconcept of (C, D) then it is a more specific concept containing less objects that have more attributes in common. The set of all formal concepts of a context $\mathbb{K}$ together with the subconcept relation is called the **concept lattice** of $\mathbb{K}$ and it is denoted with $\mathfrak{B}(\mathbb{K})$. The concept lattice is in fact a complete lattice. The set of the concept extents of all formal concepts of $\mathfrak{B}(\mathbb{K})$ is denoted with $\mathfrak{B}_1(\mathbb{K})$ and the set of all concept intents is denoted with $\mathfrak{B}_2(\mathbb{K})$. The family of sets $\mathfrak{B}_1(\mathbb{K})$ is a closure system on G and the family $\mathfrak{B}_2(\mathbb{K})$ is a closures system on M: A (set-theoretic) **closure system** $\mathcal{S}$ on a space Ω is a family $\mathcal{S} \subseteq 2^\Omega$ of subsets of Ω that contains the space Ω and is closed under arbitrary intersections. If a family $\mathcal{F}$ of subsets of a space Ω is not a closure system, one can compute its **closure** $\mathrm{cl}(\mathcal{F}) := \bigcap\{\mathcal{S} \mid \mathcal{S} \supseteq \mathcal{F} \ \& \ \mathcal{S} \text{ is a closure system on } \Omega\}$ that is the smallest closure system containing all sets of $\mathcal{F}$.

Every closure system $\mathcal{S}$ on Ω can be described by all valid formal implications of $\mathcal{S}$: A **formal implication** is a pair (Y, Z) of subsets of Ω, also denoted by $Y \longrightarrow Z$. We say that an implication $Y \longrightarrow Z$ is **valid** in a family $\mathcal{S}$ of subsets of Ω (which needs not to be a closure system) if every set of $\mathcal{S}$ that contains all elements of Y also contains all elements of Z. In this case we also say that the family $\mathcal{S}$ **respects** the implication $Y \longrightarrow Z$. Similarly, we say that a subset of Ω respects an implication $Y \longrightarrow Z$ if it either is not a superset of Y or if it is a superset of Z. The first component of a formal implication

is also called the **premise** or the **antecedent** and the second component is also called **conclusion** or the **consequent** of the formal implication. A formal implication is called **simple** if its premise is a singleton.

A closure system $\mathcal{S}$ can be characterized by formal implications as follows: Define for $\mathcal{S}$ the set $\mathfrak{I}(\mathcal{S})$ of all formal implications that are valid in $\mathcal{S}$. Then, given the set $\mathfrak{I}(\mathcal{S})$, the closure system $\mathcal{S}$ can be rediscovered from $\mathfrak{I}(\mathcal{S})$ as the set of all subsets of Ω that respect all formal implications of $\mathfrak{I}(\mathcal{S})$. The set $\mathfrak{I}(\mathcal{S})$ of all valid implications of a closure system $\mathcal{S}$ is usually very large. To efficiently describe a closure system, it suffices to look at a so-called implication base of $\mathfrak{I}(\mathcal{S})$: Given an arbitrary set $\mathfrak{I}$ of formal implications, we say that a further set $\mathfrak{J}$ of implications is a **base** of $\mathfrak{I}$, if we have

$$\forall M \subseteq \Omega: \quad M \text{ respects all implications of } \mathfrak{I} \iff M \text{ respects all implications of } \mathfrak{J} \qquad (1)$$

and if furthermore $\mathfrak{J}$ is minimal w.r.t. this property, i.e. for every other subset $\mathfrak{J}' \subsetneq \mathfrak{J}$ the equivalence (1) is not valid anymore. In the sequel, we will mainly deal with formal implications of the closure system of the concept intents of a given formal context $\mathbb{K}$. Such implications are sometimes also called attribute-implications to indicate that one is speaking about implications between attributes and not between objects of a context. Here, we will always use the short term implications and will also say that an implication is valid in a context $\mathbb{K}$ instead of saying that an implication is valid in the closure system of all concept intents of $\mathbb{K}$.

In the context of statistical data analysis one often has data that are not binary but for example categorical with more than two possible values. To analyze such data with methods of formal concept analysis one can use the technique of **conceptual scaling** (cf. [Ganter and Wille, 2012, p.36-45]) to fit the categorical data into a binary setting: For a categorical variable with the possible values in $\{1, \ldots, K\}$ one can introduce the K attributes "$= 1$", ...,"$= K$" and say that an object g has attribute "$= i$" if the value of K equals i. In a similar way, for an ordinal variable with possible values $\{1 < 2 < \ldots < K\}$ we can introduce the attributes "≤ 1","≤ 2",..., "$\leq K$" and say that object g has attribute "$\leq i$" if the value of object g is lower than or equal to i. One can also additionally introduce the attributes "≥ 1", ..., "$\geq K$". This concrete way of conceptually scaling an ordinal variable is called **interordinal scaling** and will be used in one example of application given in Section 6.2.

3 Detecting stochastic dominance

We now turn to the development of a technique for detecting stochastic dominance for poset-valued random variables based on linear programming and the upset-characterization of stochastic dominance.

3.1 Characterizing stochastic dominance via linear programming

Let $(V,\leq) = (\{v_1,\ldots,v_k\},\leq)$ be a finite poset[7], let $x = (x_1,\ldots,x_{n_x})$ be an i.i.d. sample of a random variable X and let $y = (y_1,\ldots,y_{n_y})$ be an i.i.d. sample of Y. Let $w^x = (w_1^x,\ldots,w_k^x)$ where w_i^x denotes the number of observed samples of X with value v_i, divided by n_x. Analogously, let $w^y = (w_1^y,\ldots,w_k^y)$ where w_i^y denotes the number of samples of Y with value v_i, divided by n_y. With $\mathcal{U}((V,\leq))$ we denote the set of all upsets of $(V,\leq)$. We identify an upset $M \in (V,\leq)$ with its characteristic vector $m \in \{0,1\}^k$ via $m_i = 1 \iff v_i \in M$. Additionally, we also identify the relation $\leq$ with the relation $\{(i,j) \mid i,j \in \{1,\ldots,k\}, v_i \leq v_j\}$, the same for the covering relation $\lessdot$. To the samples x and y we associate the empirical analogue $\hat{P}$ of the true law P via $\hat{P}(X = v_i) = w_i^x$ and $\hat{P}(Y = v_i) = w_i^y$. To check if $X \leq_{\hat{SD}} Y$ we have to check

$$\forall M \in \mathcal{U}((V,\leq)) : \hat{P}(X \in M) \leq \hat{P}(Y \in M). \tag{2}$$

Obviously, $\hat{P}(X \in M) = \langle w^x, m\rangle$ and $\hat{P}(X \in M) = \langle w^x, m\rangle$, so (2) is equivalently characterizable as

$$\begin{aligned}
&\forall M \in \mathcal{U}((V,\leq)): && \hat{P}(X \in M) \leq \hat{P}(Y \in M)\\
\iff &\forall M \in \mathcal{U}((V,\leq)): && \langle w^x, m\rangle \leq \langle w^y, m\rangle\\
\iff &\forall M \in \mathcal{U}((V,\leq)): && \langle w^x, m\rangle - \langle w^y, m\rangle \leq 0\\
\iff &\forall M \in \mathcal{U}((V,\leq)): && \langle w^x - w^y, m\rangle \leq 0\\
\iff &\sup_{M\in\mathcal{U}((V,\leq))} && \langle w^x - w^y, m\rangle \leq 0.
\end{aligned}$$

This means that the problem is characterizable as a linear program over the family $\mathcal{S} := \mathcal{U}((V,\leq))$ of subsets of V. To solve this program we can look at the concrete structure of the family $\mathcal{S}$. The family $\mathcal{S}$ consists of all upsets of $(V,\leq)$, i.e., of all sets M satisfying

$$\forall i,j \in \{1,\ldots,k\}: \quad v_i \in M \ \&\ v_i \leq v_j \Longrightarrow v_j \in M$$

which is equivalent to

$$\forall i,j \in \{1,\ldots,k\}: \quad v_i \leq v_j \Longrightarrow m_j \geq m_i$$

and this set of inequalities can be easily implemented in a linear program:

We have $X \leq_{\hat{SD}} Y$ if and only if the linear binary program

$$\begin{aligned}
\langle w^x - w^y, m\rangle &\longrightarrow \max\\
w.r.t.&\\
m &\in \{0,1\}^k\\
\forall (i,j) \in \leq: \quad m_j &\geq m_i
\end{aligned} \tag{3}$$

[7]This is actually no restriction because we are in the first place interested in detecting stochastic dominance for samples that are always finite.

has a maximal value of zero. (Note that the maximal value of (3) is always at least 0, because for $M = \emptyset$ we have $\langle w^x - w^y, (0, \ldots, 0)\rangle = 0$.) If one analyzes this above binary program further (see the last paragraph of Section 4.1 at page 21), one sees that it is not necessary to take the m_i as binary variables, one can relax the integrality conditions and solve instead the far more simple classical linear program

$$\begin{aligned} \langle w^x - w^y, m\rangle &\longrightarrow \max \\ w.r.t. & \\ m &\in [0,1]^k \\ \forall (i,j) \in \leq: \quad m_j &\geq m_i \end{aligned} \tag{4}$$

which could be further simplified to

$$\begin{aligned} \langle w^x - w^y, m\rangle &\longrightarrow \max \\ w.r.t. & \\ m &\in [0,1]^k \\ \forall (i,j) \in \lessdot: \quad m_j &\geq m_i. \end{aligned} \tag{5}$$

In the sequel we will denote the maximal value of (5) with D^+ and the optimal value one would get if one would replace maximization by minimization in (5) with D^-.

3.2 Some analysis of the linear programming approach for detecting stochastic dominance

The obtained linear program for checking dominance involves k decision variables and $|\lessdot|+k$ inequalities, where $|\lessdot|$ can be shown to be bounded by $\lfloor \frac{k}{2} \rfloor \cdot \lceil \frac{k}{2} \rceil$, which indicates that the linear program is practically manageable for real data sets. One interesting question in this context is how the feasible set of the linear program looks like in special situations and what for example the simplex-algorithm would actually do. In applied situations, the poset $(V, \leq)$ is often of the form $V = \mathbb{R}^d$ or $\{0, \ldots, K\}^d$ and $x \leq y \iff \forall i \in \{1, \ldots, d\} : x_i \leq y_i$. For checking stochastic dominance only the actually observed $x \in V$ are of interest.This helps in reducing the effective size of the poset V but at the same times makes the structure of V only implicitly given. Thus, a general analysis seems to be difficult and we therefore restrict the analysis in Section 3.2.1 to some simple examples.

3.2.1 Some examples

In this section we exemplarily discuss some examples for posets $(V, \leq)$ of the form $V = \{0, \ldots, K\}^d$ and $x \leq y \iff \forall i \in \{1, \ldots, d\} : x_i \leq y_i$. We start with the simplest example

where $d = 1$ which corresponds to a linearly ordered set $V = \{0 < 1 < \ldots < K\}$. Then the linear program (5) would translate to

$$\begin{aligned} \langle w^x - w^y, m\rangle &\longrightarrow \max \\ w.r.t. & \\ m &\in [0,1]^k \\ \underbrace{\begin{pmatrix} 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & -1 & 0 & \ldots & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix}}_{=:A} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} &\leq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \end{aligned}$$

In this case the extreme points of the feasible set are simply the vectors of the form $m^l = (0, \ldots, \underbrace{0}_{\text{l-th entry}}, 1, 1, \ldots)$, where $l \in \{0, \ldots, k\}$. For $l, l' \in \{1, \ldots, k-1\}$ it is easy to see that every two different extreme points m^l and $m^{l'}$ are adjacent because A has full rank and for m^l the inequality constraint associated to the $l-th$ row of A is strict where the other inequalities are actually equalities and to "switch" from m^l to $m^{l'}$ one simply has to switch the $l'-th$ variable from basis to non-basis and the $l-th$ variable form non-basic to basic. A similar argumentation shows that also for arbitrary $l, l' \in \{0, \ldots, k\}$ every two different extreme points are adjacent which means that applying the simplex algorithm would in this case exactly mean that one scans every extreme point, i.e. every upset, so the simplex algorithm is not better than a naive inspection of every upset. However in the case of a linearly ordered set the number of upsets is $|V|$ and thus no problem from a computational point of view.

Now we come to the more difficult cases of $d > 1$. In these situations the feasible set of the linear program appears to be not so easily describable, there seems to be no simple rule that says which extreme points are adjacent. Table 1 gives lower and upper bounds[8] for the size u of the closure system of all upsets of $\{0, \ldots, K\}^d$ for different values of K and d. One can see that for high enough K or d the closure system is very big and explicitly checking all upsets is clearly not applicable. Compared to this, in Table 2 one can see the

[8] The upper bounds were computed with the help of the Sauer-Shelah lemma ([Sauer, 1972, Shelah, 1972]). The Sauer-Shelah lemma is also closely related to Vapnik-Chervonenkis theory which we use in Section 5.2, see also Bottou [2013] or `http://leon.bottou.org/_media/papers/vapnik-symposium-2011.pdf` for the curious history of the Sauer-Shelah lemma. The lower bounds were obtained be noting that for every $l \in \{1, \ldots, K\}$ the set $A_l := \{x \in \{0, 1, \ldots, K\}^d \mid \sum_{i=1}^{d} x_i = l\}$ of all K-bounded multisets of rank l is an antichain and thus for every non-empty set $B \subseteq A_l$ we get a different upset $\uparrow B$. Thus, $u \geq \sum_{l=1}^{K} (2^{|A_l|} - 1) + 2$, where the last $+2$ comes from noting that also the empty set and the whole set V are upsets, and the cardinality $|A_l|$ can be computed recursively.

number of iterations a dual simplex algorithm did take to get a solution. (For the objective function we simply took a standard normally distributed sample.) This indicates that with the linear programming approach the problem is still solvable for larger values of K and d.

		d							
		1	2	3	4	5	6	7	8
K=1	lower bound	2	5	16	95	2110	1.1e+06	6.9e+10	1.2e+21
	upper bound	2	10	9.2e+01	1.5e+04	1.1e+08	3.4e+16	5.1e+31	1.5e+64
K=2	lower bound	3	15	2.7e+02	6.6e+05	2.3e+15	2.8e+42	2.0e+118	
	upper bound	3	129	1.3+06	2.1e+18	1.4e+53	1.5e+154		
K=3	lower bound	4	37	1.0e+04	2.0e+13	9.1e+46	4.0e+174		
	upper bound	4	2516	4.2e+12	8.6e+49	4.6e+187			
K=4	lower bound	5	83	1.1e+06	4.1e+25	4.9e+114			
	upper bound	5	68405	1.6e+22	4.7e+106				
K=5	lower bound	6	177	3.4e+08	9.2e+43	1.3e+235			
	upper bound	6	2391495	2.1e+34	5.5e+196				
K=6	lower bound	7	367	2.9e+11	3.5e+69				
	upper bound	7	102022809	7.0e+49					
K=7	lower bound	8	749	7.1e+14	3.6e+103				
	upper bound	8	5130659560	1.0e+68					
K=8	lower bound	9	1515	4.9e+18	1.6e+147				
	upper bound	9	296881218693	6.9e+89					

Table 1: Upper and lower bounds for the size u of the closure system of all upsets of $\{1, \ldots, K\}^d$ for different values of K and d.

	d						
K	1	2	3	4	5	6	7
1	0	0	7	18	18	92	239
2	4	3	19	156	796	3861	23002
3	3	78	208	1901	4456	24628	27271
4	17	86	626	3518	23002	24173	24923
5	12	200	2380	10987			
6	29	353	2023	23002			
7	60	396	4959				
8	87	572	7698				

Table 2: Number of iterations for solving the linear program via dual simplex for detecting stochastic dominance for $V = \{0, \ldots, K\}^d$ for different values of K and d. The objective function was a standard normally distributed random sample.

3.2.2 Duality

In this section, we analyze the dual linear program of program (4) for detecting first order stochastic dominance. The most interesting inside will be that this dual program can be interpreted as a special kind of mass transportation problem.

In order to determine the dual program of program (4), first note that the second class of constraints of problem (4) can equivalently be rewritten as

$$\forall i, j \in \{1, \ldots, k\} : m_i \geq I_{ij} \cdot m_j \tag{6}$$

where $I_{ij} := \mathbb{1}_{<}((v_j, v_i))$ and $<$ denotes the strict part of the partial order $\leq$. By defining for each $i \in \{1, \ldots, k\}$, the matrix $M^{(i)} \in \mathbb{R}^{k \times k}$ via

$$M^{(i)}_{pq} = \begin{cases} I_{ip} & \text{if } p = q \\ -1 & \text{if } q = i \wedge q \neq p \\ 0 & \text{else} \end{cases} \tag{7}$$

one then can reformulate the linear programming problem (4) by the equivalent linear programming problem

$$\begin{aligned} \langle w^x - w^y, m \rangle &\longrightarrow \max \\ w.r.t. & \\ m_1, \ldots, m_k &\geq 0 \\ \begin{pmatrix} E_k \\ M^{(1)} \\ \vdots \\ M^{(k)} \end{pmatrix} \cdot m &\leq (\underbrace{1, \ldots, 1}_{k-times}, \underbrace{0, \ldots, 0}_{k^2-times})^T =: b \end{aligned} \tag{8}$$

where E_k denotes the k-dimensional unit matrix. Define $w^{xy} := w^x - w^y$ and $z := (x_1, \ldots, x_k, z_{11}, \ldots, z_{1k}, \ldots, z_{k1}, \ldots, z_{kk})$ and let b be defined as in the constraints of the above linear program (8). Then the dual linear program of (8) is given by:

$$\begin{aligned} \sum_{l=1}^{k} x_l = \langle b, z \rangle &\longrightarrow \min \\ w.r.t. & \\ z \in \mathbb{R}^{k+k^2}_{\geq 0} & \\ \begin{pmatrix} E_k & M^{(1)T} & \ldots & M^{(k)T} \end{pmatrix} \cdot z &\geq \begin{pmatrix} w^{xy}_1 \\ \vdots \\ w^{xy}_k \end{pmatrix} \end{aligned} \tag{9}$$

In order to investigate what duality theory can teach us about our original problem, we

rewrite the program (9) as:

$$\sum_{l=1}^{k} x_l \longrightarrow \min \tag{10}$$

$$w.r.t.$$

$$z \in {\mathbb{R}_+}^{k+k^2}$$

$$\forall i \in \{1,\ldots,k\} : x_i - \sum_{s\in\{1,\ldots,k\}\setminus\{i\}} z_{is} + \sum_{s\in\{1,\ldots,k\}\setminus\{i\}} I_{si}\cdot z_{si} \geq w_i^{xy}$$

For variables z_{is} with $I_{is} = 0$, for finding an optimal solution one can always set z_{is} to zero, because such z_{is} are not present in the objective function and do occur separated only in the $i-th$ inequality constraint with a negative sign. Thus, the program can be simplified to

$$\forall i \in \{1,\ldots,k\} : x_i - \sum_{s\in\{1,\ldots,k\}\setminus\{i\}} I_{is}\cdot z_{is} + \sum_{s\in\{1,\ldots,k\}\setminus\{i\}} I_{si}\cdot z_{si} \geq w_i^{xy},$$

which again can be simplified to

$$\forall i \in \{1,\ldots,k\} : x_i - \sum_{\{s:v_s<v_i\}} z_{is} + \sum_{\{s:v_i<v_s\}} z_{si} \geq w_i^{xy}. \tag{11}$$

Note that the resulting program (10) with the rewritten version (11) of the constraints is very similar, yet not identical to the mass transport algorithm for detecting stochastic dominance discussed in [Range and Østerdal, 2013, p. 5]: In case the optimal objective of the program equals 0, the values z_{ij}^* can be interpreted as the probability masses that need to be transported from strictly greater elements to strictly smaller elements w.r.t. $\leq$ in order to obtain the distribution of X from the distribution of Y (which exactly corresponds to characterization iii) of first order stochastic dominance that was recalled in the introduction). The main difference of our program (10) and the problem discussed in [Range and Østerdal, 2013, p. 5] is that, while there the authors have two classes of constraints, one class for the masses transported into each node and one class for the masses transported out of each node, our set of constraints considers the masses that are transported inside in- and out of each node simultaneously.

Note that there are also attempts to interpret the value of the sum $\sum_{ij} z_{ij}^*$ of the optimal z_{ij}^* values, or a weighted version of it in cardinal settings (see [Tarp and Østerdal, 2007, p.19-20]), as a measure for the extent of stochastic dominance that is given in the situation under consideration. However, as discussed in further detail in Section 3.3, in this paper we argue that in order to detect the extent of stochastic dominance using the optimal value of (10) might be a more sensible indicator for the extent of stochastic dominance since it avoids certain counter-intuitive characteristics.

In order to get a better impression of the structure of the above dual programming problem, we consider the following example: Suppose the poset V consists of seven elements, namely $V = \{v_1, \ldots, v_7\}$. Moreover, suppose the partial order $\leq$ is specified by the following incidence matrix M:

$$M = \begin{array}{c|ccccccc} & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 \\ \hline v_1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ v_2 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ v_3 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ v_4 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ v_5 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ v_6 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ v_7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

where we have that $M_{ij} = 1$ if and only if $v_i \leq v_j$. Finally, suppose we have observed samples of X and Y and computed the vectors w^x and w^y. Then, the dual linear programming problem from (10) takes the following form:

$$\begin{aligned} \sum_{l=1}^{7} x_l &\longrightarrow \min \\ w.r.t. & \\ (x_1, \ldots, x_7, z_{11}, z_{12}, \ldots, z_{76}, z_{77}) &\in {\mathbb{R}_+}^{56} \\ x_1 + (z_{21} + z_{31} + z_{41} + z_{51} + z_{61} + z_{71}) &\geq w_1^{xy} \\ x_2 - (z_{21}) + (z_{52} + z_{62} + z_{72}) &\geq w_2^{xy} \\ x_3 - (z_{31}) + (z_{53} + z_{63} + z_{73}) &\geq w_3^{xy} \\ x_4 - (z_{41}) + (z_{54} + z_{64} + z_{74}) &\geq w_4^{xy} \\ x_5 - (z_{51} + z_{52} + z_{53} + z_{54}) + (z_{75}) &\geq w_5^{xy} \\ x_6 - (z_{61} + z_{62} + z_{63} + z_{64}) + (z_{76}) &\geq w_6^{xy} \\ x_7 - (z_{71} + z_{72} + z_{73} + z_{74} + z_{75} + z_{76}) &\geq w_7^{xy} \end{aligned}$$

First, consider the observed samples lead to vectors $w^x = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$ and $w^y = (\frac{1}{7}, 0, 0, 0, \frac{2}{7}, \frac{2}{7}, \frac{2}{7})$. For that case, the optimal objective of the above programming problem is 0 (which, due to duality and Proposition 3.1, also indicates that Y first-order stochastic dominates X). A corresponding optimal solution vector is given by $(x_1^*, \ldots, x_7^*, z_{11}^*, \ldots, z_{77}^*)$, where every component equals 0 except of $z_{54}^* = z_{63}^* = z_{72}^* = \frac{1}{7}$. As discussed before, the z_{ij}^* variables exactly describe how the distribution of X can be obtained from the distribution of Y by a finite number of probability mass transfers to strictly smaller elements with respect to the partial order $\leq$. In our example, the distribution of X can be obtained from that of Y by transferring mass $\frac{1}{7}$ from node v_5 to v_4, mass $\frac{1}{7}$ from node v_6 to v_3 and mass $\frac{1}{7}$ from node v_7 to v_2. This is illustrated in Figure 1.

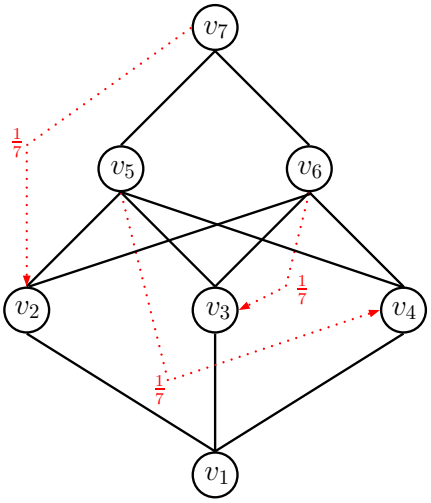

Figure 1: Mass transfer problem for $w^x = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$ and $w^y = (\frac{1}{7}, 0, 0, 0, \frac{2}{7}, \frac{2}{7}, \frac{2}{7})$.

A natural question is the following: Do optimal solutions of the programming problem (10) still possess a meaningful interpretation for the case that Y does not stochastically dominate X? To address this question, suppose we, instead of the previous situation, observed the vectors $w^x = \frac{1}{28} \cdot (4, 5, 6, 2, 1, 3, 7)$ and $w^y = \frac{1}{28} \cdot (4, 2, 5, 7, 6, 1, 3)$. In that case, the optimal objective of our example is $\frac{6}{28}$ (which indicates that Y does not stochastically dominate X by the same argument as given above) and an optimal solution vector is given by $(x_1^*, \ldots, x_7^*, z_{11}^*, \ldots, z_{77}^*)$, where all components equal 0 except $x_6^* = \frac{2}{28}$, $x_7^* = \frac{4}{28}$, $z_{52}^* = \frac{3}{28}$ and $z_{53}^* = \frac{1}{28}$. Indeed, also in the case of a non-dominant Y we receive a straightforward interpretation: Compared to the case of stochastic dominance, where the whole probability mass can be transported from higher values to lower values to obtain X from Y, in the case of non-dominance, not all mass can be transported and the optimal value of (10) could be understood as the amount of probability mass that cannot be transported and thus has to be externally introduced to supply X with enough probability mass. Again, the optimal solution is illustrated in Figure 2.

3.3 The minimal value as a measure of the extent and the argmin as an insight into the actual manifestation of dominance

With the linear program (5) we can detect stochastic dominance. However, as already betoken, generally one is not only interested in the presence or absence of stochastic dominance, one would also like to get some rough idea about the "extent" of dominance. In our very general setting of random variables/data with only a partially ordered scale of measurement, a reasonable definition of the term *extent of dominance* is not straight forward. Therefore, we will firstly go one step back and reconstruct, how the upset characterization of stochastic dominance, that was introduced only in purely

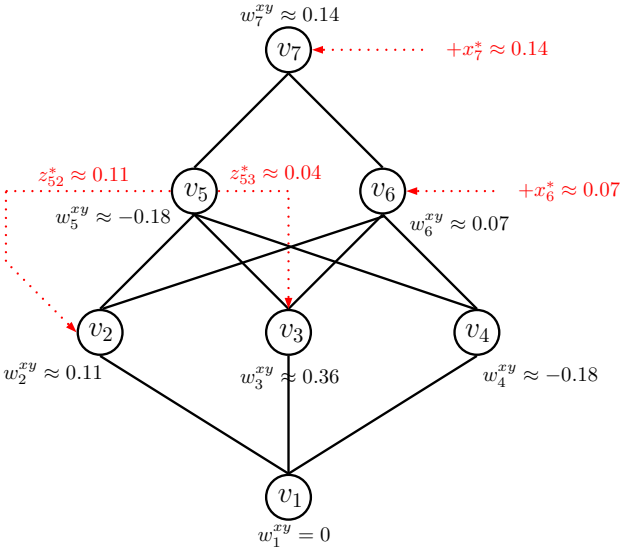

Figure 2: Mass transfer problem for $w^x = \frac{1}{28} \cdot (4, 5, 6, 2, 1, 3, 7)$ and $w^y = \frac{1}{28} \cdot (4, 2, 5, 7, 6, 1, 3)$.

mathematical terms until now, can be concretely interpreted in conceptual terms. We will do this by relying on one prototypic example of poverty/inequality analysis[9]. To make it simple, we will start with the notion of income poverty as a simple example of univariate poverty/inequality analysis. Consider for example that one is interested in the differences of income-poverty in two countries. One simple approach is here to firstly define a so-called poverty line c and to say that every person with income below the poverty line c can be termed *poor* whereas all persons with an income above the poverty line c can be termed *non-poor*. (The terms *poor* and *non-poor* are meant here in a purely descriptive sense free from value judgment). If the poverty line could be defined in a reasonable manner from a substance matter point of view, then for "measuring" the extent of inequality, one can compare the proportions of the *poor* persons in the two countries (also called head count ratio), for example by looking at the differences of the proportions in the wo countries. If it is difficult to specify the poverty line c, then one can get rid of the need for the specification of the poverty line by simultaneously looking at every reasonable poverty line c. If, independently from the choice of the poverty line c, the proportion of the poor is always greater for one country than for the other country, one can reasonably say that the income-poverty in one country is clearly greater than the income-poverty in the other country, which is exactly saying that one country is dominated by the other w.r.t. classical univariate first order dominance. In the situation of a given, fixed poverty line c, one can measure the extent of poverty for example with the income gap ratio, which is the relative difference between the income of the poor and the poverty line. The difference of the income gap ratios can then be used to measure

[9] Of course, in other concrete situations, the conceptual reconstruction done here could be less convincing.

the difference in the extent of poverty between the two countries. Compared to this, in the situation of multivariate inequality analysis, the involved dimensions of poverty like e.g. health or education are often only of ordinal scale of measurement. Of course, other dimensions like income have a higher scale of measurement, but this does not help in assessing, which amount of increase of income can compensate for which decrease in health or education. Of course, one can standardize every dimension in a reasonable way, but this would lead to a relative notion of inequality. Here, we go another way and use a notion of "extent" of dominance that is not related to units of the different dimensions but that is only based on the proportion of persons that are termed *non-poor* (or *poor*).

To do so, let us firstly think about the translation of the notion of a poverty line to the multivariate case: In the univariate case of income inequality we said that persons with income below the poverty line c could be termed *poor*, and the persons with income above the poverty line could be termed *non-poor*. In the multivariate setting, the way to term persons as *poor* or *non-poor* is only restricted by the underlying partial order $\leq$. If one terms one person i as *poor*, then one should also declare a person j as *poor* if the attributes x_j of person j are all lower than or equal to the attributes x_j of person x_j (i.e. $x_i \leq x_j$). This is exactly the concept of a downset of a partially ordered set: Every downset M of a poset $(V, \leq)$ is a reasonable concretion of the term *poor* in the sense that all $x \in M$ can be called *poor* and all $x \notin M$ could be called *non-poor*. The notion of a downset is the natural generalization of the notion of a poverty line to the multivariate case. In the sequel, we will deal with upsets instead of downsets. Dually to the notion of downstes, the notion of upsets[10] models the reasonable concretions of the term *non-poor* instead of the term *poor*. We can now interpret the maximal value D^+ and the minimal value D^- of the linear program (5) for detecting stochastic dominance: For the prototypic example of inequality analysis, if the maximal value D^+ is zero, then we know that X is stochastically dominated by Y, meaning that the proportion of the non-poor persons in subpopulation Y is greater than or equal to the proportion of the non-poor persons in subpopulation X, independently from the concretion of the term *non-poor*. Furthermore, in this case the minimal value D^- can be interpreted as some measure of the extent of stochastic dominance. The value $D^-$is exactly the difference between the proportions of non-poor persons of the two countries for that concretion of the term *non-poor* that is the most conservative in the sense that it maximizes the absolute value of the difference in proportions between the two countries. The extent of stochastic dominance can thus be measured to some extent with the minimal value of (5) with the following clear interpretation: The absolute value of D^- is exactly the proportion of poor persons in the poorer country that would have to be made non-poor to make the proportions of the poor (and thus also he proportions of the non-poor) in both countries the same, where the notion of *poor* is the most conservative, i.e., for every other reasonable notion of *poor* one would only have to make a smaller proportion of poor people of the poor country non-poor to make the proportions of the poor the same in each country.

[10]Note that the complements of downsets are upsets and that the complements of upsets are downsets.

As already mentioned, in [Tarp and Østerdal, 2007, p.19-20], a similar quantity for measuring the extent of dominance was proposed. There, the authors use the characterizing property (iii) of stochastic dominance and they measure the extent of dominance by the (weighted) amount of probability mass that is needed to obtain the density f_X from f_Y. However, the used measure seems to be not very sensitive in the following specific sense: Assume for simplicity real-valued random variables X and Y. Assume further that X is a simple transformation of Y, concretely $X := Y - \varepsilon$ with positive but very small ε. Then, one would have to transport the whole probability mass to obtain f_X from f_Y no matter how small the value ε is. This seems to be a very undesirable property of this measure of the extent of dominance. In such a situation, the measure D^- of extent proposed here behaves differently. For example if X and Y are normally distributed, then the maximal value D^- of (5) would be strictly increasing in ε and would furthermore converge to zero if ε converges to zero. For constant random variables X and Y the measure D^- would be either zero (if ε is zero) or one (if ε is grater than zero). This seems counter-intuitive at first glance, the measure is insensitive to the distance between X and Y. Actually this behavior is adequate, because one presumes only an ordinal scale of measurement for X and Y here, and thus one cannot reasonably measure the distance between X and Y. The fact that one could actually be sensitive to the distance between X and Y in the normally distributed case is due to the fact, that with our measure, we do not directly measure (non-existing) distances in the space of the values of X and Y, instead we measure indirectly the "distance" between X and Y by the amount of probability mass that has to be transported to compensate inequality for the most conservative choice of a poverty line.

While the value of D^- gives a quantitative insight in the extent of dominance, the upset U, where D^- is attained additionally gives a further more qualitative insight into how the worst possible concretion of the term *poor*, for which the maximal inequality in poverty is attained, looks like. This could be interesting for example if one is interested in the question, if the purely mathematical formalization of *poor* and *non-poor* via upsets is maybe too rigorous and if an extreme value of the test statistic is only due to an upset representing a very "skewed" concretion of the term *non-poor*, that could maybe be excluded as a reasonable concretion of the term *poor* because of substance matter considerations. In such a situation on may use the regularization techniques developed in Section 5.3.

3.4 Checking stochastic dominance as a linear program on a closure system

One important point to note is that the way we incorporated the property of being an upset was by introducing simple inequalities of the form $m_j \geq m_i$ for all pairs $(i, j) \in \preccurlyeq$. In the language of formal implications this means that we demanded that an upset should contain with every v_i also every $v_j \succcurlyeq v_i$, which is exactly saying that the formal implication

$\{v_i\} \longrightarrow \{v_j\}$ should be valid in the closure system of upsets. In fact, for the closure system $\mathcal{S}$ of upsets, the essential implications are the implications of the form $\{v_i\} \longrightarrow \{v_j\}$ with $v_i \lessdot v_j$, because all these implications are respected by $\mathcal{S}$ and they already describe $\mathcal{S}$ in the sense that they are a base of all valid implications. (Obviously there are further redundant implications like e.g. $\{v_i\} \longrightarrow \{v_i\}$ or $\{v_i, v_k\} \longrightarrow \{v_j\}$ with $v_i \leq v_j$ and v_k arbitrary.) For the case of upsets we were especially lucky, because all such essential implications had a simple premise (meaning that the premise A in $A \longrightarrow B$ is a singleton) and thus we could implement this implications via simple inequalities $m_i \leq m_j$ and could furthermore drop the integrality constraints. There are other situations that are such simple, too: Due to Birkhoffs theorem ([Birkhoff, 1937]), every (finite) closure system that is additionally closed under union[11] is describable via simple formal implications, and examples of such kinds of closure systems arise for example in the context of quasi-ordinal knowledge space theory (see, e.g., [Doignon and Falmagne, 2012, p.38-40], note also that there are neat connections between knowledge space theory and formal concept analysis, cf., [Rusch and Wille, 1996]). A natural question is now: Can we still solve the problem of maximizing/minimizing a linear function on an arbitrary closure system that is not describable via simple implications and could this have some application? The answer is simply yes: The next sections will give two examples of closure systems that are either explicitly given by an implication base or that are implicitly given as the concept extents of a given formal context.

4 Linear programming on general closure systems

4.1 The case of closure systems efficiently described by formal implications

In some situations, a closure system that is very big can still be efficiently described by an implication base of all valid implications. One example is the closure system $\mathcal{C}(\mathbb{R}^2)$ of all convex sets in $\mathbb{R}^2$ that could be of interest in the context of spatial statistics. The set of all valid implications of $\mathcal{C}(\mathbb{R}^2)$ is given by $\mathfrak{I} = \{A \longrightarrow \mathrm{co}(A) \mid A \subseteq \mathbb{R}^2\}$ where co is the operator that maps a set to its convex hull. Because of Carathéodory's theorem for convex hulls[12] the system $\mathcal{L} = \{A \longrightarrow \mathrm{co}(A) \mid A \subseteq \mathbb{R}^2, |A| \leq 3, \mathrm{co}(A) \supsetneq A\}$ is an implication base of $\mathfrak{I}$. In statistical applications in the context of spatial data analysis, for example in ecology, one is interested in differences in the spatial distribution of different species, for example male and female Pacific cods in the eastern Bering Sea analyzed in Syrjala [1996]. To describe differences in the spatial distributions of the two subpopulations, one can use common test statistics[13]: Here, test statistics of many different statistical tests are

[11] The closure system of upsets is such a system.

[12] Carathéodory's theorem states that if a point $x \in \mathbb{R}^d$ lies in the convex hull of a set P of points, then there exists a subset $Q \subseteq P$ of at most $d+1$ points such that x lies also in the convex hull of Q.

[13] Here, in the first place we are mainly interested in the test statistic as a descriptive tool, the problem of inference will be discussed in a general setting in Section 5.

available, for example generalizations of the Kolmogorv-Smirnov test or generalizations of the Cramér von Mises test (see [Syrjala, 1996]) could be used. For the Kolmogorov-Smirnov type generalization one determines for every rectangular area the difference in the proportion of male and female cods. Then one computes the maximal difference over all rectangular areas. This method needs a specification of a rectangular coordinate system and the results are dependent on the concrete choice of this coordinate system. Opposed to this, one could also simply look not only at all rectangular, but instead at all convex areas and then compute the maximal difference. This would be exactly an optimization of a linear function on a closure system. The result of the optimization on all convex sets instead of all rectangular areas would then be independent of the choice of a coordinate system, because for the definition of convexity, no specification of a coordinate system is needed at all. If one did observe cods at altogether k spatial points $(v_1, \ldots, v_k)$ then one actually does not need to look at the whole closure system $C(\mathbb{R}^2)$, it suffices to look only at the projected closure system $\mathcal{C}(\{v_1, \ldots, v_k\}) := \{A \cap \{v_1, \ldots, v_k\} \mid A \in \mathcal{C}(\mathbb{R}^2)\}$. To compute the test statistic one can solve a binary program, where all implications are implemented as inequality constraints. This method is generally applicable for arbitrary closure systems with a given implication base: For a given implication base $\mathcal{L}$ of an arbitrary finite closure system, one can compute the statistic

$$\sup_{A \subseteq V,\ A \text{ respects } \mathcal{L}} \langle w^x - w^y, \mathbb{1}_A \rangle$$

by solving the following binary program:

$$\begin{aligned} \langle w^x - w^y, m \rangle &\longrightarrow \max \\ w.r.t. & \\ m &\in \{0,1\}^k \\ \forall (Y, Z) \in \mathcal{L} : \sum_{i: v_i \in Y} m_i - \frac{1}{|Z|} \sum_{i: v_i \in Z} m_i &\leq |Y| - 1. \end{aligned}$$

Here, for any given implication $Y \longrightarrow Z$ of $\mathcal{L}$, the corresponding inequality constraint of the binary program is automatically satisfied if the premise of $Y \longrightarrow Z$ is not fulfilled, because then the left hand side is lower than $|Y| - 1$. If the premise Y is fulfilled, then the corresponding inequality translates to $-\frac{1}{Z} \sum_{i: v_i \in Z} m_i \leq -1$ or equivalently to $\frac{1}{Z} \sum_{i: v_i \in Z} m_i \geq 1$, thus demanding that all m_i with $i \in Z$ should be one, meaning that if the set described by the indicator function $(m_1, \ldots, m_k)$ contains all elements of Y, then it should also contain all elements of Z. In our concrete situation of convex sets we would have to solve a binary program with n decision variables and $\mathcal{O}(n^3)$ inequality constraints. Unfortunately, generally, the integrality constraints cannot be dropped, here. Thus, the program becomes very difficult to solve if n is large.

An important case where one can actually drop the integrality constraints is the case where one has only simple implications: In this case one can implement every simple implication $\{v_i\} \longrightarrow \{v_j\}$ as the inequality $m_i \leq m_j$. To see that one can drop the integrality constraints in this case, observe that any feasible vector $(m_1, \ldots, m_k)$ of the relaxed program with some $m_l \notin \{0, 1\}$ is not an extreme point of the feasible set of the relaxed program, since for $\varepsilon > 0$ chosen small enough (for example $\varepsilon = 1/2 \min\{|m_i - m_l| \mid i \in \{1, \ldots, n\}, m_i \neq m_l\}$) it can be represented as the convex combination of the two feasible vectors $m + x^\varepsilon$ and $m - x^\varepsilon$ where $x^\varepsilon \in \mathbb{R}^k$ is defined as $x_i^\varepsilon = \begin{cases} \varepsilon & \text{if } m_i = m_l \\ 0 & \text{else} \end{cases}$.

4.2 The case of closure systems efficiently described by a generating formal context

Closure systems also naturally arise in the theory of formal concept analysis: the family of all formal concept extents (as well as the family of all concept intents) of a concept lattice is a closure system. Furthermore, every arbitrary closure system can be represented as a closure system of extents (or intents) of an appropriately chosen formal context[14]. In statistical applications, it appears natural to take as objects the observed data points, for example persons in a social survey. As attributes one can take the values of different variables of interest, for example the answers of the persons to different questions. (If the questions are yes-no questions, then they can be incorporated directly, otherwise one can apply the method conceptual scaling to get binary data, cf. Section 2.2.) The formal concept lattice then gives valuable qualitative information about different subgroups of persons that supplied response patterns that belong to the same formal concept and thus share specific attributes. If one is interested in differences between different subgroups (e.g., male and female participants) w.r.t. the answers to the questions, one could look at every formal concept and analyze the differences between the subpopulations that belong to the given concept. Often, the concept lattice is very large and it becomes difficult to look at every formal concept. Then, one can look for example only on that concepts, for which the difference between the proportions of persons belonging to this concept in each subgroup is maximal or minimal. This is exactly the problem of maximizing a linear function on the closure system of concept extents. If the whole concept lattice can be computed explicitly, then one can simply explicitly compute for every extent the difference in proportions between both subpopulations. However, in many situations the concept lattice is so big that it is very hard to compute all extents/intents explicitly to perform the optimization. (In the worst case, a formal context can have $\min(2^{|G|}, 2^{|M|})$ associated formal concepts.) In this situation one can use the fact that a pair (A, B) with $A \subseteq G$ and

[14]For a closure system $\mathcal{S} \subseteq 2^V$ take the formal context $\mathbb{K} := (V, \mathcal{S}, \in)$, then the formal extents are exactly the sets of $\mathcal{S}$. Analogously, for the dual context $\mathbb{K} := (\mathcal{S}, V, \ni)$, the formal intents are exactly the sets of $\mathcal{S}$.

$B \subseteq M$ is a formal concept iff

$$A = B' \ \& \ B = A'$$

or equivalently iff

$$\forall g \in G, m \in M : \mathbb{1}_A(g) = \min_{m \in B} \mathbb{1}_{m'}(g) \ \& \ \mathbb{1}_B(m) = \min_{g \in A} \mathbb{1}_{g'}(m). \tag{12}$$

Characterization (12) can be used to describe the property of being a formal concept with the help of the following characterizing inequalities:

$$\forall g \in G, m \in B : \mathbb{1}_A(g) \leq \mathbb{1}_{m'}(g) \tag{13}$$

$$\forall g \in A, m \in M : \mathbb{1}_B(m) \leq \mathbb{1}_{g'}(m) \tag{14}$$

$$\forall g \in A, m \in B : \mathbb{1}_A(g) \geq \sum_{m \in B} \mathbb{1}_{m'}(g) - |B| + 1 \quad \& \tag{15}$$

$$\mathbb{1}_B(m) \geq \sum_{g \in A} \mathbb{1}_{g'}(m) - |A| + 1. \tag{16}$$

Equations (13) and (21) capture the fact that $\mathbb{1}_A(g) \leq \min_{m \in B} \mathbb{1}_{m'}(g)$ and $\mathbb{1}_B(m) \leq \min_{g \in A} \mathbb{1}_{g'}(m)$, respectively. Equations (15) and (16) say that $\mathbb{1}_A(g) \geq \min_{m \in B} \mathbb{1}_{m'}(g)$ and $\mathbb{1}_B(m) \geq \min_{g \in G} \mathbb{1}_{g'}(m)$, respectively, which is equivalent to the condition that if an object g has all attributes of B then it has to be in the extent A and that if an attribute m is shared by all objects of A, then it should be in the intent B. The characterization via inequality constraints can be used to optimize a linear function of the indicator function of the extents (or the intents, or both) with a binary program: Let $G = \{g_1, \ldots, g_m\}$ be the set of objects, $M = \{m_1, \ldots, m_n\}$ the set of attributes and let $A \in \{0,1\}^{m \times n}$ be a matrix describing the incidence I with the interpretation $A_{ij} = 1 \iff$ object number i has attribute number j. A formal concept can then be described by a binary vector $z = (z_1, \ldots, z_m, z_{m+1}, \ldots, z_{m+n}) \in \{0,1\}^{m+n}$, where the first m entries describe the extent via $z_i = 1$ iff object i belongs to the extent and the last n entries describe the intent as $z_{j+m} = 1$ iff attribute j belongs to the intent. The characterizing constraints (13) - (16) would then translate to the conditions

$$\forall (i,j) \text{ s.t. } A_{ij} = 0 : \qquad z_i \leq 1 - z_{j+m} \quad \& \quad z_{j+m} \leq 1 - z_i \tag{17}$$

$$\forall i \in \{1, \ldots, m\} : \qquad z_i \geq \sum_{k : A_{ik} = 1} z_{k+m} - \sum_{k=1,\ldots,n} z_{k+m} + 1 \tag{18}$$

$$\forall j \in \{1, \ldots, n\} : \qquad z_{j+m} \geq \sum_{k : A_{kj} = 1} z_k - \sum_{k=1,\ldots,m} z_k + 1. \tag{19}$$

$$\tag{20}$$

have to be satisfied. This could be simplified to the condition

$$\forall (i,j) \text{ s.t. } A_{ij} = 0 : \qquad z_i \leq 1 - z_{j+m} \tag{21}$$

$$\forall i \in \{1, \ldots, m\} : \qquad \sum_{k:A_{ik}=0} z_{k+m} \geq 1 - z_i \tag{22}$$

$$\forall j \in \{1, \ldots, n\} : \qquad \sum_{k:A_{kj}=0} z_k \geq 1 - z_{j+m}, \tag{23}$$

which has the following intuitive interpretation:

For every 0-entry in the i-th row and the j-th column of the matrix A we have:

1. if $A_{ij} = 0$ and if object g_i belongs to the extent, then necessarily attribute m_j cannot belong to the intent and vice versa.
2. If object g_i does not belong to the extent, then there exists at least one attribute m_k of the intent, that the object g_i does not have.
3. Dualy, if attribute m_j does not belong to the intent, then there exists at least one object g_k of the extent, that has not attribute m_j.

Thus, we can compute the maximum

$$\max_{(A,B)\in\mathfrak{B}(\mathbb{K})} \langle w^{ext}, \mathbb{1}_A \rangle + \langle w^{int}, \mathbb{1}_B \rangle$$

of an arbitrary linear objective function $(w_1^{ext}, \ldots, w_n^{ext}, w_1^{int}, \ldots, w_n^{int})$ of both the extents and the intents by solving the binary program

$$\begin{aligned} \langle (w_1^{ext}, \ldots, w_m^{ext}, w_1^{int}, \ldots, w_n^{int}), (z_1, \ldots, z_m, z_{m+1}, \ldots, z_{m+n}) \rangle &\longrightarrow \max \\ w.r.t. & \\ \forall (i,j) \text{ s.t. } A_{ij} = 0 : z_i &\leq 1 - z_{j+m} \\ \forall i \in \{1, \ldots, m\} : \sum_{k:A_{ik}=0} z_{k+m} &\geq 1 - z_i \\ \forall j \in \{1, \ldots, n\} : \sum_{k:A_{kj}=0} z_k &\geq 1 - z_{j+m} \end{aligned} \tag{24}$$

All in all, we would thus have to solve a **binary** program with $m + n$ variables and $|\{(i,j \mid A_{ij} = 0)\}| + m + n$ constraints. This problem can become cumbersome if the formal context is big enough, especially because one cannot simply drop the integrality-constraints. However, in practical applications, often only the number of objects is large and the number of items is medium-sized. If one further analyzes the binary program, then one observes that the inequalities concerning the objects and the constraints concerning the attributes

are separated in the sense that if one relaxes only the integrality constraints of the variables $z_1, \ldots, z_m$ describing the extent, then the optimum of the associated relaxed mixed binary program is still (also) attained at a binary solution and thus one can relax the integrality constraints of the extent. The reason is that for fixed and binary $(z_{m+1}, \ldots, z_{m+n})$ the inequalities (21) and (22) are either redundant or reduce to equality constraints of the form $z_i = 0$ or $z_i = 1$ and inequality (23) is either redundant or demands that a sum of z_k's associated with the extent is greater or equal to 1. If one of the z_k's is not binary, than at least one other $z_{k'}$ has to be greater than zero. This means that for an appropriately chosen[15] $\varepsilon > 0$ the vectors $(z_1, \ldots, z_k + \varepsilon, \ldots, z_{k'} - \varepsilon, \ldots, z_{m+n})$ and $(z_1, \ldots, z_k - \varepsilon, \ldots, z_{k'} + \varepsilon, \ldots, z_{m+n})$ are still feasible with respect to the relaxed feasible set and this shows that non-integer points are no extreme-points of the restricted polytope where the binary variables describing the intent are fixed. Thus, the optimal value for the relaxed program is always also attained at a binary solution.

5 Statistical inference

We now treat the question of inference. Coming back to the example of detecting stochastic dominance, we were able to detect stochastic dominance in a sample. The natural question of inference is now: What can we reasonably infer about stochastic dominance in the population we sampled from? From a substance matter perspective, one would supposedly be interested for example in the hypotheses

$$\begin{aligned} H_0: &\quad X \text{ is not stochastically dominated by } Y \quad \text{vs} \\ H_1: &\quad X \text{ is stochastically dominated by } Y. \end{aligned}$$

However, a reasonable consistent classical statistical test of this pair of hypotheses is not reachable since already in the univariate case where the distribution function characterizes stochastic dominance, we have the problem that for every $X \leq_{SD} Y$, in every arbitrarily small neighbourhood[16] of Y we can find some $\tilde{Y}$ with $X \not\leq_{SD} \tilde{Y}$. To circumvent this problem, one can modify the hypotheses, for example by switching the roles of H_0 and H_1 (for consistent statistical tests of this kind in the univariate case, see, e.g., [Barrett and Donald, 2003]). Here, we go a slightly different way. Since the value of D^+ characterizes $X \leq_{SD} Y$ via $X \leq_{SD} Y$ iff $D^+ = 0$ and D^- characterizes $Y \leq_{SD} X$ via $Y \leq_{SD} X$ iff $D^- = 0$ and furthermore $X \nsim_{SD} Y$ (where $X \nsim_{SD} Y$ means $X \not\leq_{SD} Y$ & $Y \not\leq_{SD} X$) iff $D^+ > 0$ & $D^- < 0$) we can simply test, if D^+ and D^- are significantly different from zero. (In the case of for example D^+ is significantly positive and D^- is not significantly negative,

[15] The choice of ε depends on all inequalities that involve z_k and $z_{k'}$ but since there are only finite many constraints, ε can in fact be chosen small enough and still greater than zero.

[16] This is meant w.r.t. e.g., the Kolmogorov-Smirnov distance. Note also, that in our situation, we have not much freedom of choice of other distances that induce other neighborhood concepts, since we can only make use of the partially ordered scale of measurement of X and Y.

we cannot directly conclude $X \leq_{SD} Y$ but at least $Y \not\leq_{SD} X$ and the possibility $X \not\leq_{SD} Y$ is only possible due to upsets A with $P(X \in A) \geq P(Y \in A)$ where the difference $P(X \in A) - P(Y \in A)$ is only slightly positive.) In the sequel we will put focus on D^+ and also do not explicitly correct for multiple testing if considering both D^+ and D^- simultaneously. Actually, conceptually, here we do not take the inference problem as the primitive and do not rigorously test a beforehand exactly stated hypothesis by doing a statistical test that provides us with a descriptively interpretable test statistic as a by-product. Instead, we see it a little bit the other way around: In the first place, we would like to get a good, conceptually rigorous descriptive insight into the data by not relying on traditional approaches based on somehow "arbitrarily chosen" location measures[17] summarizing the data by one number and then comparing the obtained numbers. Instead, by relying on stochastic dominance, we in a sense somehow look simultaneously at all reasonable location measures and if we know $X \leq_{SD} Y$, then we also know that every reasonable location measure[18] would give a lower (or equal) number to X than to Y. This is a conceptually much more reliable statement than simply comparing numbers (of course with the drawback of being less decisive). Only in a second step we think in statistical terms about to which extent the conceptually rigorous statement of stochastic dominance can be translated from the sample to the population.

5.1 Permutation-based tests

Now, let us come to the purely statistical aspects of inference for detecting stochastic dominance. (All considerations are similarly valid for linear optimization on general closure systems.) In the simple univariate case of real-valued, continuously distributed random variables X, Y, for the two-sample case under $H_0 : F_X = F_Y$, the distribution of the test statistic D^+ (and also D^- and $D := \max\{D^+, -D^-\}$) is independent of the true law F_X, has known asymptotics and can be furthermore computed exactly for identical sample sizes (see, e.g., [Pratt and Gibbons, 2012, Chapter 7]). Opposed to this, in the general multivariate situation, the statistic D^+ is not distribution free, anymore: Firstly, the distribution of D^+ depends on the concrete structure of the poset $(V, \leq)$: If the relation $\leq$ is very sparse, then the set of all upsets is very large and one would generally expect that D^+ would typically have higher values than for the case of a very dense relation $\leq$. Secondly, also the interplay between the structure of $(V, \leq)$ and the unknown true law is also of relevance: For example in a very large poset $(V, \leq)$ with a very sparse relation $\leq$ it could be still the case that the most probability mass is living on a much smaller subset $W \subseteq V$ on which the restricted relation $\leq \cap\, W \times W$ is actually very dense. This suggests that a rigorous analytic treatment of the distribution of D^+ seems to be only partially

[17] Note the non-classical scale of measurement we are dealing with, here.

[18] One of the few location measures that does not respect first order stochastic dominance is the mode. But note that the mode appears most naturally if we have a categorical or an interval scale of measurement, the mode seems to give no valuable information if we want to analyze inequality which is a genuinely ordinal concept.

possible[19]. Thus, a natural alternative is to apply a two sample observation-randomization test (permutation test, see, e.g., [Pratt and Gibbons, 2012, Chaper 6]), here. The procedure for evaluating the distribution of D^+ under $H_0 : P_X = P_Y$, which is the least favorable case of $\tilde{H}_0 : \quad D^+_{true} := \sup_{A \in \mathcal{U}((V,\leq))} P_X(A) - P_Y(A) = 0 \quad (\iff X \leq_{SD} Y))$ is straightforward:

1. Let a sample $x = (x_1, \ldots, x_{n_x})$ of size n_x for subpopulation X and a sample $y = (y_1, \ldots, y_{n_y})$ of size y of subpopulation Y be given.
2. Compute the statistic D^+ for the actually observed data.
3. Take the pooled sample $z = (x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_y})$.
4. Take all index sets $I \subseteq \{1, \ldots, n_x + n_y\}$ of size n_x and compute the test statistic D^+_I that would be obtained for a virtual sample $\tilde{x} = (z_i)_{i \in I}$ for population X and $\tilde{y} = (z_i)_{i \in \{1,\ldots,n_x+n_y\} \setminus I}$ for subpopulation Y.
5. Order all D^+_I in increasing order
6. Reject H_0 if the test statistic D^+ for the actually observed data is greater than the $\lceil \gamma \cdot |I| \rceil$-th value of the increasingly ordered values D^+_I, where γ is the envisaged confidence level.

In step 4 one has to compute the test statistic for a very huge number of resamples, thus one usually does not compute the test statistic for all resamples but only for a smaller number of randomly chosen resamples. In the context of linear programming on closure systems, the computation of the test statistic for one resample could be already computational demanding for very complex data sets, so the application of observation-randomization tests has some limitations, here.

5.2 Conservative bounds via Vapnik-Chervonenkis theory

Beyond applying resampling schemes for inference there is the further possibility to apply Vapnik-Chervonenkis theory (see, e.g., [Vapnik and Kotz, 1982]) to obtain conservative bounds for the test statistic: In Vapnik-Chervonenkis theory, among other things, one analyzes the distribution of

$$\sup_{A \in \mathcal{S}} |P_n(A) - P(A)|$$

or

$$\sup_{A \in \mathcal{S}} |P_n(A) - P'_n(A)|,$$

[19] Actually, there exists some literature on the asymptotic distribution of the optimal value of a random linear program (e.g., [Babbar, 1955, Sengupta et al., 1963, Prèkopa, 1966]). However, this literature seems to be not applicable in our situation, because in our case, under the null hypothesis, the random objective function is symmetrically distributed around the zero vector, such that the assumption of a unique optimal basis for the asymptotic linear program (cf. [Prèkopa, 1966, Theorem 5]) is not satisfied.

where P is an unknown probability law and P_n is the empirical law associated with an i.i.d.-sample of size n (and P'_n is the empirical law associated to a further independently drawn sample of the same size n). Here, the family $\mathcal{S}$ can be any arbitrary family of subsets of a given space Ω. In our situation, the family $\mathcal{S}$ is the underlying closure system of interest. The Vapnik-Chervonenkis inequalities (cf., [Vapnik and Kotz, 1982, p.170-172]) then state that

$$P\left(\sup_{A\in\mathcal{S}}|P_n(A)-P(A)|>\varepsilon\right)\leq 6\, m^{\mathcal{S}}(2n)\ e^{-n\varepsilon^2/4} \quad \text{and} \tag{25}$$

$$P\left(\sup_{A\in\mathcal{S}}|P_n(A)-P'_n(A)|>\varepsilon\right)\leq 3\, m^{\mathcal{S}}(2n)\ e^{-n\varepsilon^2}. \tag{26}$$

These inequalities[20] can be used to get conservative critical values for a one sample and a two sample test. (In the sequel, we will put focus on the two sample situation.) The crucial quantity involved in the right hand sides of these inequalities is the so-called **growth function**

$$\begin{aligned} m^{\mathcal{S}}(k) :=& \max_{A\subseteq\Omega,|A|=k}\Delta_{\mathcal{S}}(A), \text{ where} \\ \Delta_{\mathcal{S}}(A) :=& |\{S\cap A \mid S\in\mathcal{S}\}| \end{aligned}$$

describes the cardinality of the projection of the family $\mathcal{S}$ on the set A. Obviously, if $\mathcal{S}$ is finite, then the growth function $m^{\mathcal{S}}(k)$ is always lower than or equal to the cardinality of $\mathcal{S}$ and thus $|\mathcal{S}|$ can be used to get a bound for the left hand sides of (25) and (26). Actually, in our setting, we will use as the underlying space always the subset V_{ess} of all actually observed values of the basic set V and an associated closure system $\mathcal{S}\subseteq 2^{V_{ess}}$ on the restricted space $\Omega := V_{ess}$. (Note that the projection of a closure system $\mathcal{S}'\subseteq 2^{\Omega'}$ on Ω' onto a subset $\Omega\subseteq\Omega'$ via $\mathcal{S}'_{|\Omega}=\{S\cap\Omega \mid S\in\mathcal{S}'\}$ is again a closure system on Ω.) Thus, with $A=V_{ess}$ we have $\Delta_{\mathcal{S}}(A)=|\mathcal{S}|$ and $m^{\mathcal{S}}(2n)=|\mathcal{S}|$, such that the bound $|\mathcal{S}|$ is a sharp bound for $m^{\mathcal{S}}(2n)$. If the family $\mathcal{S}$ is explicitly given, we could thus work with the computable bound $|\mathcal{S}|$. The far more interesting situation appears if the family $\mathcal{S}$ is very large and only implicitly given. Then there is another important bound (see [Vapnik and Kotz, 1982, p.167]) on the growth function that is related to the **Vapnik-Chervonenkis dimension (V.C.-dimension)** of the family[21] $\mathcal{S}$:

$$m^{\mathcal{S}}(k)\leq 1.5\frac{k^{VC-1}}{(VC-1)!},$$

[20] There is a bunch of similar Vapnik-Chervonenkis type inequalities that could be of help here, see, e.g., the summary given in Table 1 of [Vayatis and Azencott, 1999, p.4].

[21] Originally, Vapnik-Chervonenkis theory was mainly developed to be able to deal with infinite families $\mathcal{S}$. Here, we have finite families $\mathcal{S}$, and if we would know $|\mathcal{S}|$ then, in our setting, we would better bound $m^{\mathcal{S}}(k)$ by $|\mathcal{S}|$ instead of using the Vapnik-Chervonenkis dimension since in our setting of $\mathcal{S}\subseteq 2^{V_{ess}}$, this dimension essentially only provides an upper bound for the cardinality of $|\mathcal{S}|$. If the family $\mathcal{S}$ is very large and is only implicitly given, then the V.C.-dimension can still provide a good computable bound for $m^{\mathcal{S}}(k)$. Note further that sometimes also other bounds for $m^{\mathcal{S}}(2n)$ can be useful, for example for finite Ω wa have $m^{\mathcal{S}}(2n)\leq 2^{|\Omega|}$.

where VC is the Vapnik-Chervonenkis dimension of the family $\mathcal{S}$, that is defined as the cardinality of the largest possible subset A that can be shattered by $\mathcal{S}$: One says that a set A can be **shattered** by $\mathcal{S}$ (or alternatively that A is shatterable w.r.t. $\mathcal{S}$) if the projection of $\mathcal{S}$ on A contains all subsets of A, i.e. $2^A = \{S \cap A \mid S \in \mathcal{S}\}$ or equivalently $\Delta_{\mathcal{S}}(A) = 2^{|A|}$. In many cases, the V.C.-dimension cannot be computed explicitly. However, in our context it shows up that we can compute the V.C.-dimension either with the help of binary programs or with a sharp characterization of the V.C.-dimension. Of course, the Vapnik-Chervonenkis inequalities provide only very conservative bounds for inference. (Note that the right hand sights of (25) and (26) do not depend on the true law P.) If one is able to perform an observation randomization test, then one should do it instead of dealing with the conservative Vapnik-Chervonenkis inequalities. However, the Vapnik-Chervonenkis inequalities give us some guidance for dealing with situations where the closure system is so big that one would expect that the distribution of the test statistic is behaving too ugly to allow for a sensible statistical test with enough power. In such a situation, we can use Vapnik-Chervonenkis theory to appropriately reduce the closure system to hope for making the tail distribution of the test statistic more well-behaved[22]. If one appropriately reduces the cardinality of the closure system, then one could hope for a test statistic that has a better power for the detection of a "systematic" deviation[23] from H_0. This possibly increased power would then come along with a smaller and thus less fine-grained closure system $\mathcal{S}$ that is then not so sensitive to very specific alternatives. Note that the V.C.-inequality is essentially based on the effective size of $\mathcal{S}$. Thus, if $\mathcal{S}$ is explicitly given, one can simply drop some sets of $\mathcal{S}$ to make $\mathcal{S}$ smaller. However, in our situation, we often have a closure system that is implicitly given and only nicely describable because it is a closure system. A simple removal of some sets of $\mathcal{S}$ is thus not possible because sets of $\mathcal{S}$ are not explicitly given and an arbitrary removal of some sets could lead to a family $\mathcal{S}'$, that is not a closure system, and thus not easily describable, anymore. The beauty of V.C.-theory in this situation lies in the fact that if we can compute the V.C.-dimension by supplying a shatterable set A of maximal cardinality, then we also have a straightforward possibility to tame $\mathcal{S}$: Since big shatterable sets A make $\mathcal{S}$ very big, we can drop some or all elements of such sets A to tame $\mathcal{S}$ efficiently. Actually, one would not completely remove A (or a subset of A) for the whole data analysis, but only for the construction of the closure system under which the final data analysis takes place. Before explaining how this is exactly meant in different situations and how we ensure that the tamed system is still a closure system (cf. Section 5.3), we will now firstly characterize shatterable sets and the V.C.-dimension for different closure systems in the next section.

[22] Of course, V.C.-theory gives us only bounds on the tail behavior of the test statistic and no direct insight into the actual behavior of the tails, so a sharpening of the bound does not necessarily mean that the actual tail behavior will be getting better if we reduce the V.C.-dimension of the closure system.

[23] Of course, one cannot hope for a more powerful statistic w.r.t. every thinkable deviation from H_0 but only for a better power for detecting deviations that are not "too complex" w.r.t. V.C.-dimension.

5.2.1 The Vapnik-Chervonenkis dimension of several special closure systems

This section is actually very technical and not really necessary to understand the basic ideas in the following sections. The reader more interested in the basic concepts can thus skip this section and Section 5.2.2. For the reader interested in Vapnik-Chervonenkis theory and the reader interested in a detailed understanding, we would like to recommend the following sections, because, though looking a bit technical, there are no deep or cumbersome ideas involved in the following theorems. Contrarily, the relation between Vapnik-Chervonenkis theory and formal concept analysis seems to be very natural. Maybe somehow surprising, there seems to be not too much research that directly connects formal concept analysis and Vapnik-Chervonenkis theory, the only works in this direction, the authors are aware of, are the papers Anthony et al. [1990a,b], Albano and Chornomaz [2017], Chornomaz [2015], Albano [2017a,b], Makhalova and Kuznetsov [2017].

Definition & Proposition 1. *Let $\mathcal{S} \subseteq 2^\Omega$ be a closure system on Ω. Let furthermore $\mathfrak{I}(\mathcal{S})$ be the set of all formal implications the closure system $\mathcal{S}$ respects. A set $M \subseteq \Omega$ is called* **implication-free** *if there is no formal implication $(A, B) \in \mathfrak{I}$ where A and B are disjoint non-empty subsets of M. A set M is shatterable w.r.t. $\mathcal{S}$ if and only if it is implication-free and thus the Vapnik-Chervonenkis dimension of $\mathcal{S}$ is the maximal cardinality of an implication-free set $M \subseteq \Omega$.*

Definition 5 (Vapnik-Chervonenkis principal dimension (VCPI/VCPF))**.** *Let $(V, \leq)$ be a partially ordered sett. The* **Vapnik-Chervonenkis principal ideal dimension** *(VCPI) is the Vapnik-Chervonenkis dimension of the family*

$$\mathfrak{pi}((V, \leq)) := \{\downarrow x \mid x \in V\} = \{\{y \mid y \leq x\} \mid x \in V\}\}$$

of all principal ideals of $(V, \leq)$. If $(V, \leq)$ is a complete lattice, then $\mathfrak{pi}((V, \leq))$ is a closure system. In this case we also say that a set $M \subseteq V$ is **join-shatterable** *if it is shatterable w.r.t. the family $\mathfrak{pi}((V, \leq))$. Analogously, the* **Vapnik-Chervonenkis principal filter dimension** *(VCPF) is the Vapnik-Chervonenkis dimension of the family*

$$\mathfrak{pf}((V, \leq)) := \{\uparrow x \mid x \in V\} = \{\{y \mid y \geq x\} \mid x \in V\}\}$$

of all principal filters of $(V, \leq)$. If $(V, \leq)$ is a complete lattice, then $\mathfrak{pf}((V, \leq))$ is a closure system. In this case we also say that a set $M \subseteq V$ is **meet-shatterable** *if it is shatterable w.r.t. the family $\mathfrak{pf}((V, \leq))$.*

Theorem 1 (Motivating the notions *join-shatterable* and *meet-shatterable*)**.** *A subset M of a complete lattice $(V, \leq)$ is join-shatterable if and and only if we have for every $x \in M$:*

$$x \not\leq \bigvee M \backslash \{x\}. \tag{27}$$

Analogously, a subset M of a complete lattice $(V, \leq)$ is meet-shatterable if and and only if we have for every $x \in M$:

$$x \not\geq \bigwedge M \backslash \{x\}. \tag{28}$$

Proof. We only proof the first statement, the second statement can be proofed analogously.
if: Let $B \subseteq M$. Take $A :=\downarrow \bigvee B \in \mathfrak{pi}((V,\leq))$. Then $A \cap M \supseteq B$ since $\forall b \in B : b \leq \bigvee B$. Additionally, for every $x \in M\backslash B$ we have $B \subseteq M\backslash\{x\}$ and thus $x \notin A$, since if $x \in A$, because of $A \subseteq\downarrow \bigvee M\backslash\{x\}$ we would get $x \leq \bigvee M\backslash\{x\}$ which would be a contradiction to (27). Thus $A \cap M = B$ and because B was an arbitrary subset of M, we can conclude that M is shatterable.
only if: Let $x \leq \bigvee M\backslash\{x\}$ for some $x \in M$. Then the set $M\backslash\{x\}$ is not shatterable w.r.t. $\mathfrak{pi}((V,\leq))$, because every $a \in V$ with $\forall y \in M\backslash\{x\} : y \leq a$ is an upper bound of $M\backslash\{x\}$ and thus $a \geq \bigvee M\backslash\{x\} \geq x$. But this means, that every set $A =\downarrow a \in \mathfrak{pi}((V,\leq))$ that contains all elements of $M\backslash\{x\}$ necessarily also contains x which shows that in fact $M\backslash\{x\}$ is not shatterable. □

Theorem 2. *For every finite join-shatterable set M of a finite*[24] *complete lattice $(V,\leq)$ there exists another join-shatterable set J_M of join-irreducible elements of $(V,\leq)$ that has the same cardinality as M. This means that for determining the Vapnik-Chervonenkis principal ideal dimension it is enough to look at join-shatterable sets of join-irreducible elements.*

Proof. Let $M \subseteq V$ be a finite shatterable set. If all elements of M are join-irreducible then we are done. If there exists an $x \in M$ that is not join-irreducible we can find a join-irreducible element z such that the set $\tilde{M} := M\backslash\{x\} \cup \{z\}$ is still join shatterable. Since M is assumed to be finite, we can replace step by step every join-reducible element of M by a join-irreducible element and thus obtain a shatterable set of join-irreducible elements with the same cardinality: So let $x \in M\backslash\mathcal{J}(V)$. Then $x = \bigvee B$ for some set $B \subseteq \mathcal{J}(V)$. Furthermore, we have $z \not\leq \bigvee M\backslash\{x\}$ for at least one $z \in B$, because otherwise we would have $\bigvee M\backslash\{x\} \geq \bigvee B = x$ which is in contradiction with the assumption that the set M is join-shatterable. Now, take $\tilde{M} := M\backslash\{x\} \cup \{z\}$. Then, $\tilde{M}$ is join-shatterable. To see this, observe that M and $\tilde{M}$ only differ in the elements x and z and $z \leq x$. Thus $z \not\leq \bigvee M\backslash\{x\} = \bigvee \tilde{M}\backslash\{z\}$ and for every other $y \in \tilde{M}$ we have $y \not\leq \bigvee \tilde{M}\backslash\{y\}$ because otherwise we would have $y \leq \bigvee \tilde{M}\backslash\{y\} \leq \bigvee M\backslash\{y\}$ which is in contradiction with M being join-shatterable. □

Theorem 3. *The Vapnik-Chervonenkis principal ideal dimension VCPI (and also the Vapnik-Chervonenkis principal filter dimension VCPF) of a poset $(V,\leq)$ is bounded by its order dimension*[25] $\mathbf{odim}((V,\leq))$.

Proof. Let $d := \mathbf{odim}((V,\leq))$ and let $L_1,\ldots,L_d$ be d linear orders representing $\leq$ via $x \leq y \iff \forall i \in \{1,\ldots,d\} : xL_iy$. We show that every set M of more than d elements

[24] The finiteness assumption can be dropped if one only assumes that every element $x \in V$ can be written as a supremum of join-irreducible elements of V. This is for example the case if there are no infinite descending chains in V.

[25] Remember that the order dimension of a poset $(V,\leq)$ is the smallest number d of linear orders $L_1,\ldots,L_d \subseteq V \times V$ such that the relation $\leq$ can be represented as the intersection of these linear orders via $x \leq y \iff \forall i \in \{1,\ldots,d\} : xL_iy$.

of V is not join-shatterable: Take M with $|M| > d$ and take for every $i \in \{1, \ldots, d\}$ that element $x_i \in M$ that is the greatest element of M w.r.t. the linear order L_i. Then every principal ideal $\downarrow a$ that contains all the x_i necessarily also contains every further element $y \in M \backslash \{x_1, \ldots, x_d\} \neq \emptyset$ because for every $i \in \{1, \ldots, d\}$ we have $yL_ix_iL_ia$. □

Theorem 4. *Let $\mathbb{V} = (V, \leq)$ be a finite complete lattice. If $\mathbb{V}$ is distributive*[26]*, which can be characterized by saying that the condition*

$$\forall B \subseteq \mathcal{J}(\mathbb{V}) \forall x \in \mathcal{J}(\mathbb{V}) : \quad x \leq \bigvee B \Longrightarrow x \leq z \text{ for some } z \in B \tag{29}$$

is fulfilled, then the Vapnik-Chervonenkis principal dimension of $\mathbb{V}$ is exactly the width of $\mathcal{J}(\mathbb{V})$ and because of Birkhoffs theorem we have $\mathbb{V} \cong (\mathcal{D}(\mathcal{J}(\mathbb{V}))$ and the width of $(\mathcal{J}(\mathbb{V}))$ is exactly the order dimension of $\mathbb{V}$, so in this case we have $VCPI(\mathbb{V}) = \mathbf{odim}(\mathbb{V})$.

Proof. Because of Theorem 2 we only have to look at the set $\mathcal{J}(\mathbb{V})$ of the join-irreducible elements of $\mathbb{V}$. Let d denote the width of $\mathcal{J}(\mathbb{V})$. It is clear that a join-shatterable set $M \subseteq (\mathcal{J}(\mathbb{V}))$ necessarily is an antichain. Thus VCPI is lower than or equal to d. To see that $VCPI = d$ take an antichain A of size d. Then this antichain is obviously shatterable because for all $x \in A$ we have $x \not\leq \bigvee A \backslash \{x\}$ since if $x \leq \bigvee A \backslash \{x\}$ because of (29) we would have $x \leq z$ for some z in $A \backslash \{x\}$, but this would be in contradiction with A being an antichain. □

Definition & Proposition 2 (Vapnik-Chervonenkis upset dimension: Simply the width)**.** *Let $\mathbb{V} = (V, \leq)$ be a poset and $\mathcal{U}(\mathbb{V})$ be the set of all upsets of $\mathbb{V}$. Then the Vapnik-Chervonenkis dimension of $\mathcal{U}(\mathbb{V})$ is called the* **Vapnik-Chervonenkis upset dimension**. *The Vapnik-Chervonenkis upset dimension is identical to the width of $\mathbb{V}$, because the shatterable sets are exactly the implication-free sets, which are in this case the antichains of $\mathbb{V}$. Analogously, the Vapnik-Chervonenkis dimension of all downsets is also equal to the width.*

Definition 6 (Vapnik-Chervonenkis formal context dimension (VCC))**.** *Let $\mathbb{K} := (G, M, I)$ be a formal context. Let*

$$\mathcal{S} := \mathfrak{B}_1((G, M, I)) = \{A \subseteq G \mid (A, B) \in \mathfrak{B}((G, M, I)) \text{ for some } B \subseteq M\}$$

be the closure system of all concept extents. The **Vapnik-Chervonenkis formal concept dimension** *(VCC) is defined as the Vapnik-Chervonenkis dimension of $\mathcal{S}$.*

Theorem 5 (cf. also [Albano and Chornomaz, 2017, Albano, 2017a,b])**.** *Let $\mathbb{K} := (G, M, I)$ be a formal context and let $\mathcal{S} := \mathfrak{B}_1((G, M, I))$. Then a set $\{g_1, \ldots, g_l\} \subseteq G$ of objects is shatterable w.r.t. $\mathcal{S}$ if and only if there exists a set $\{m_1, \ldots, m_l\} \subseteq M$ of attributes such that*

$$\forall i, j \in \{1, \ldots, l\} : (g_i, m_j) \in I \iff i \neq j. \tag{30}$$

[26] A lattice $\mathbb{L}$ is called distributive if we have $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ for arbitrary $x, y, z \in \mathbb{L}$.

Proof. **if:** Let $A \subseteq \{g_1, \ldots, g_l\}$. Take the formal concept (A'', A'). Then A'' contains all $g_i \in A$ and for all j with $g_j \notin A$ because of $m_j \in A'$ we have $g_j \notin A''$ and thus $A'' \cap \{g_1, \ldots, g_l\} = A$ which shows that $\{g_1, \ldots, g_l\}$ is shatterable w.r.t. $\mathcal{S}$.
only if: If $\{g_1, \ldots, g_l\}$ is shatterable then for every g_i there exists a formal concept (A_i, B_i) such that $g_i \notin A_i$ and $\forall j \in \{1, \ldots, l\} \backslash \{i\} : g_i \in A_i$. But this means that for every $i \in \{1, \ldots, l\}$ there exists an attribute m_i such that $(g_i, m_i) \notin I$ and $\forall j \in \{1, \ldots, l\} \backslash \{i\} : (g_i, m_j) \in I$. □

Corollary 1. *The Vapnik-Chervonenkis formal context dimension of a context (G, M, I) is equal to the Vapnik-Chervonenkis formal context dimension of the dual context (M, G, I^∂), where $I^\partial = \{(m, g) \mid g \in G, m \in M, gIm\}$.*

5.2.2 Computation of the Vapnik-Chervonenkis dimension

In this section we shortly propose some methods to actually compute the Vapnik-Chervonenkis dimension for different closure systems.

Computing the Vapnik-Chervonenkis dimension if the closure system is given via formal implications

If the closure system $\mathcal{S}$ is given by all valid formal implications, then computing the V.C.-dimension can be done by searching for an implication-free set A of maximal cardinality. To do this, one can solve the following binary program:

$$\sum_{i=1}^{k} m_i \longrightarrow max \tag{31}$$

$$w.r.t. \tag{32}$$

$$\forall (Y, Z) \in \mathfrak{I}(\mathcal{S}) : \sum_{i:v_i \in Y} m_i + \frac{1}{|Z|} \sum_{i:v_i \in Z} m_i \leq |Y| \tag{33}$$

$$m = (m_1, \ldots, m_k) \in \{0, 1\}^k \tag{34}$$

Here, condition (33) codifies the demand that for a valid implication $Y \longrightarrow Z$ a shatterable (implication-free) set A necessarily cannot contain any element of Z if it contains all elements of Y. Instead of the whole set of implications in (33) one can also use only that valid implications $Y \longrightarrow Z$ where Y is minimal (in the sense that $\tilde{Y} \longrightarrow Z$ is not valid anymore for every $\tilde{Y} \subsetneq Y$) and Z is maximal (in th sense that $Y \longrightarrow \tilde{Z}$ is not valid anymore for every $\tilde{Z} \supsetneq Z$). This set of implications is referred to as the **generic base** in formal concept analysis (cf., e.g., [Bastide et al., 2000], where also an algorithm for extracting the generic base is given). Note that in (33) one cannot use an arbitrary implication base: For example the implication base

$$\mathfrak{I} := \{\{v_1\} \longrightarrow \{v_2\}, \{v_2, v_3\} \longrightarrow \{v_4\}\}$$

induces the further implication $\{v_1, v_3\} \longrightarrow \{v_4\}$ and thus the set $A = \{v_1, v_3, v_4\}$ is not shatterable, but the "anti-implications"

$$\mathfrak{I} := \{\{v_1\} \longrightarrow \{\neg v_2\}, \{v_2, v_3\} \longrightarrow \{\neg v_4\}\}$$

obtained by demanding (33) only for the implication base $\mathfrak{I}$ would not exclude the set A although it is not shatterable.

If one wants to compute for example the Vapnik-Chervonenkis principal ideal dimension VCPI of an explicitly given complete lattice $\mathbb{L} = (L, \leq)$, one can firstly construct the formal context $\mathbb{K} := (V, V, \geq)$. Then, the closure system of all intents of this context is exactly the set of all principal ideals of $(L, \leq)$ and one can compute the generic base of all implications that are valid in this closure system. Finally, one can build and solve the binary program (31). Actually, due to Theorem 2 it suffices to look only at the reduced context where join-irreducible elements of $\mathbb{L}$ are removed.[27]

Computing the Vapnik-Chervonenkis upset dimension: Computing the width

If one wants to compute the Vapnik-Chervonenkis upset dimension, in principle one can use the binary program (31), but since the Vapnik-Chervonenkis upset dimension is simply the width, one can also use other more efficient algorithms to compute the width. One possibility is to reformulate the problem of computing the width of a poset as a matching problem in a bipartite graph: Define the bipartite graph $G = (V \times \{1\}, V \times \{2\}, E)$ where the set of vertices is the disjoint union of V and V and the two parts of G are essentially two copies of the poset V and an edge $e = ((v, 1), (w, 2))$ is in E iff $v < w$. Now one can compute a maximal matching in G. The maximum matching then corresponds to a minimum size chain partitioning of V where two elements v and w with $v < w$ are in the same partition iff the edge $((v, 1), (w, 2))$ is in the maximal matching. The number of partitions is then $|V| - m$ where m is the size of the maximal matching. This means that we have found a minimal chain partitioning of V with size $|V| - n$ which is due to Dilworth's theorem identical to the maximal cardinality of an antichain, i.e., the width. To actually compute the maximum matching one can use e.g. the algorithm of Hopcroft and Karp (Hopcroft and Karp [1971]), which would have time complexity $\mathcal{O}(|V|^{\frac{5}{2}})$ in our situation.

Computing the Vapnik-Chervonenkis formal context dimension VCC

[27] Note that for an explicitly given poset $(V, \leq)$ that is not a complete lattice, the family $\mathfrak{pi}((V, \leq))$ of all principal ideals is generally not a closure system, but one can look at the closure system that is generated by all principal ideals of $(V, \leq)$ (of course, without the need of explicitly computing it). Then, to make the computation more efficient, one can similarly remove all reducible attributes (and also all reducible objects) from the context $\mathbb{K} = (V, V, \geq)$, where an attribute a is called reducible if the formal concept $(\{a\}', \{a\}'')$ is meet-reducible and an object o is called reducible if the formal concept $(\{o\}'', \{o\}')$ is join-reducible.

To compute the Vapnik-Chervonenkis formal context dimension VCC one can simply make use of Theorem 5. One can equivalently express condition (30) of Theorem 5 by saying that A set $A = \{g_1, \ldots, g_l\}$ of objects is shatterable w.r.t. $\mathfrak{B}_1(\mathbb{K})$ if and only if there exists a set $B = \{m_1, \ldots, m_l\}$ such that for every object $g \in A$ there exists exactly one attribute $m \in B$ with $(g, m) \notin I$ and if furthermore, for every attribute $m \in B$ there exists also exactly one object $g \in A$ with $(g, m) \notin I$. These two conditions can also be incorporated via inequality constraints. Thus, we can compute the Vapnik-Chervonenkis formal context dimension of a context $\mathbb{K}$ by jointly analyzing pairs (A, B) of an object set A and a an intent set B satisfying (30). With the notation of Section 4.2 we have to solve the binary program

$$
\begin{aligned}
& z_1 + \ldots + z_m \longrightarrow \max \\
& \qquad w.r.t. \\
\forall i \in \{1, \ldots, m\}: \quad & (n-1) \cdot z_i + \sum_{j : A_{ij} = 0} z_{j+m} \leq n && (35) \\
& -z_i + \sum_{j : A_{ij} = 0} z_{j+m} \geq 0 && (36) \\
\forall j \in \{1, \ldots, n\}: \quad & (m-1) \cdot z_{j+m} + \sum_{i : A_{ij} = 0} z_i \leq m && (37) \\
& -z_{j+m} + \sum_{i : A_{ij} = 0} z_i \geq 0. && (38)
\end{aligned}
$$

Here, the constraints (35) and (36) are redundant if z_i is zero, e.g., if object g_i does not belong to the envisaged shatterable set A. If object g_i is in the envisaged shatterable set A, then (35) demands exactly that there is maximal one attribute m_j in the associated attribute set B with $(g_i, m_j) \notin I$ and constraint (36) further demands that there is also at least one such attribute. The constraints (37) and (38) analogously codify the dual statement where the roles of objects and attributes are exchanged. Here, unfortunately one generally cannot drop any integrality constraint, so the computation of the V.C. formal context dimension is generally very hard.

5.3 Taming the monster: pruning closure systems via Vapnik-Chervonenkis theory

The last section showed how to compute the V.C.-dimension for several closure systems and how to identify shatterable sets of maximal cardinality. The ability to identify such big shatterable sets supplies us with a simple possibility of effectively taming the closure system by removing such big shatterable sets to get a test statistic that is less crude in the sense that one gets better bounds in (25) and (26) due to a lower V.C.-dimension.

Concretely, for e.g. the closure system $\mathcal{U}((V,\leq))$ of upsets, every upset $M \in \mathcal{U}((V,\leq))$ can be characterized by the set $\min(M)$ of all minimal elements of M via $M = \uparrow \min(M)$. To tame $\mathcal{U}((V,\leq))$ one can compute an antichain[28] A of maximal cardinality and then remove this antichain A (or a subset$\tilde{A}$ of A) from $\mathcal{U}((V,\leq))$ by considering not all upsets $\mathcal{U}((V,\leq)) = \{\uparrow B \mid B \subseteq V\}$ but only the family $\mathcal{S}' = \{\uparrow B \mid B \subseteq V\backslash A\}$ of all upsets that are generated by $V\backslash A$ (or $V\backslash\tilde{A}$). This family is generally not a closure system, anymore, but one can simply take not the family $\mathcal{S}$ but the closure system[29] $\tilde{\mathcal{S}}$ that is generated by $\mathcal{S}'$ via $\tilde{\mathcal{S}} := \mathrm{cl}(\mathcal{S}') = \bigcap\{\mathcal{F} \mid \mathcal{F}$ closure system on $V\backslash A,\ \mathcal{F} \supseteq \mathcal{S}'\}$.

For taming the Vapnik-Chervonenkis formal context dimension of a given formal context $\mathbb{K}$ one can similarly look for objects involved in a shatterable set of maximal cardinality and then take the closure system of the concept extents of the formal concept lattice generated by the modified context where the objects involved in a shatterable set of maximal cardinality are removed. Generally, two issues arise here:

Firstly, for a closure system $\mathcal{S}$ of V.C.-dimension VC one usually has more than one shatterable set of size VC. To effectively tame the closure system one therefore has to remove the first found shatterable set of size VC and then one has to look at further shatterable sets of size VC and remove them, too. In this situation, it could be the case that the result of the taming procedure depends upon which shatterable set of maximal cardinality was removed first. To avoid this problem, one can alternatively look jointly at all shatterable sets of maximal cardinality and remove them all. However, this could have the effect that in one step a huge number of sets is removed such that the V.C.-dimension becomes too small already in one step. Furthermore, if one decides for removing only subsets of shatterable sets, then it is not straightforward, which subsets exactly to remove and also here, the choice of the removed subsets could possibly have an impact on which set would be a shatterable set of maximal cardinality in the next step. Since the ability of removing not only whole shatterable sets but also subsets would be very helpful for taming in a very flexible way, this could be seen as a problem.

Secondly, the taming of the closure system is only a statistical "regularization procedure" that only cares for the purely statistical aspects. Thus, it is desirable to analyze the taming also with respect to its "conceptual behavior" in the sense that one should care for how flexible the tamed closure system is w.r.t. which sets are in the closure system and how fine-grained the tamed closure system thus is w.r.t a purely descriptive/conceptual point of view. This is clearly a matter of the concrete application. For the closure systems of upsets and the closure system of concept extents we will now give concrete proposals for taming that are in our view also acceptable from a conceptual point of view in the situations of the application examples given later in Sections 6.1 and 6.3.

[28] As shown in Section 5.2.1, for the closure system of all upsets of a poset $(V,\leq)$, the shatterable sets are exactly the antichains of $(V,\leq)$.

[29] For the computation of the test statistic on the tamed closure system $\tilde{\mathcal{S}}$ one does not need to compute $\tilde{\mathcal{S}}$ explicitly, see Section 5.3.3.

5.3.1 Taming upsets in the context of inequality analysis

The closure system of upsets played a crucial role in the context of stochastic dominance. One field of application of stochastic dominance is multivariate poverty or inequality analysis. In this context one does not start with a poset V, instead one has some (often totally ordered) "dimensions" of poverty/inequality. In our example of application given in Section 6.1., we have basically the 3 dimensions *Income*, *Education* and *Health*. The poset $(V, \leq)$ is then given by the three dimensional attributes of all persons in the survey equipped with the coordinate wise order (i.e. person x is poorer than or as equally poor as person y iff she is poorer than or as equally poor as y w.r.t. every dimension). Then, the concept of an upset codifies a "multivariate poverty line" in the sense that an upset M would be a reasonable concretion of the term *non-poor* by saying that every person in the set M could be termed *non-poor* and every person in the complement of M could be termed *poor*. The statement of stochastic dominance $X \leq_{SD} Y$ where X describes one subpopulation and Y another subpopulation would then mean $\forall M \in \mathcal{U}((V, \leq));\ P(X \in M) \leq P(Y \in M)$ which can be simply translated to the statement: "However the term *poor* is actually reasonably concretized, in every case the proportion of the *non-poor* persons in subpopulation corresponding to X is always lower than or equal to the proportion in the subpopulation related to Y." Now, how can we reasonable tame the closure system of upsets in this context? Since the closure system of upsets is getting very big already for small posets V, a taming by explicitly removing upsets seems hopeless, but one can use the fact that every upset M is generated by its minimal elements via $M = \uparrow \min(M)$ and look at antichains instead of upsets. One way to tame the closure system of all upsets, i.e., the closure system of all reasonable concretions of the term *non-poor* in a conceptually reasonable way could be to exclude some very "skew" concretions of the term *poor*: One can try to remove upsets generated by antichains consisting of very unbalanced elements, i.e. attributes that are very low in one dimension and at the same time very high in another dimension. To do so, one has to concretize here, what low and high means. One possibility would be to firstly standardize[30] every dimension to be $U[0, 1]$ distributed. Concretely, if $X \in \mathbb{R}^{n \times p}$ is the matrix containing the n attributes of dimension p, define for $j = 1, \ldots, p$ the univariate empirical distribution[31] function F^j according to the distribution of the j-th dimension in the sample and define $Z \in \mathbb{R}^{n \times p}$ via $Z_{ij} = F^j(X_{ij})$. Then Z is a transformation of X where every dimension $Z_{\bullet j}$ has values ranging from $\frac{1}{n}$ to 1 allowing for some kind of relative comparability of the transformed attributes with the simple interpretation that if $Z_{ij} = \frac{l}{n}$ the person i is the

[30] If one has any external substance matter insight into how some decrease in one dimension can be reasonably be compensated for by an increase in another dimension, one should try to reflect this substance matter insight into the taming procedure. Of course, the herein proposed taming procedure has to be understood as a general purpose procedure that could be substantially improved by modifications based on substance matter considerations.

[31] One can use here the complementary distribution function $F^j(x) = \frac{|\{i | X_{ij} \geq x\}|}{n}$ or the usual distribution function $F^j(x) = \frac{|\{i | X_{ij} \leq x\}|}{n}$, which would lead to identical results. We use here the complementary distribution function because it fits more to the notion of upsets.

$n-l+1$-th poorest person in the sample w.r.t. dimension i. To concretize the notion of an "imbalanced" multivariate poverty line one can firstly define a transformed multivariate attribute $Z_{i\cdot} = (Z_{i1}, \ldots, Z_{ip})$ as balanced if $\max\{Z_{i1}, \ldots, Z_{ip}\} - \min\{Z_{i1}, \ldots, Z_{ip}\} \leq \delta$ for a fixed threshold δ. Then, one can define a poverty line as balanced if it is generated by an antichain that consists only of balanced attributes. If one sets the threshold δ globally to one fixed value, then w.r.t. V.C.-dimension it can happen that the V.C.-dimension can vary drastically from region to region in the sense that e.g. for regions of medium transformed Z values there are big sets of balanced elements building an antichain whereas for extreme Z values there are only small sized antichains of balanced elements. For the statistical side of the taming procedure this could lead to a very brute taming of regions of extreme Z values without globally reducing the V.C.-dimension very much. Of course, in the proof of the Vapnik-Chervonenkis inequality one essentially deals with the cardinality of the closure system and this is actually sized down by the procedure, so the statistical taming would actually still be achieved, but only if one is taming very strongly which means that one would reduce far more sets in regions of extreme Z-values/low width than in regions of medium Z-values/high width where the density of upsets is already very high. (Note that every antichain of size k induces 2^k upsets). One can avoid this seemingly bad effect with the following localization method[32]:

First, fix some envisaged V.C.-dimension h_0. For given $\alpha \in [0,1]$ and for arbitrary $\varepsilon > 0$ define an ε-stratum around the center α as the set $M_\varepsilon(\alpha) = \{v_i \mid \forall j \in \{1, \ldots, p\} : |Z_{ij} - \alpha| \leq \varepsilon\}$. Then, for fixed α choose $\varepsilon(\alpha)$ such that the V.C.-dimension of $M_{\varepsilon(\alpha)}(\alpha)$ is lower than or equal to h_0 and such that $\varepsilon(\alpha)$ is maximal w.r.t. this property. Then collect in a set $T(h_0) := \bigcup\{M_{\varepsilon(\alpha)}(\alpha) \mid \alpha \in [0,1]\}$ all strata $M_{\varepsilon(\alpha)}(\alpha)$. The closure system $\mathcal{S}_{h_0} = \mathrm{cl}(\mathcal{F}_{h_0})$ generated by the family of sets $\mathcal{F}_{h_0} = \{\uparrow B \mid B \subseteq T(h_0)\}$ can then serve as a tamed subsystem of $\mathcal{S}$. Note that the V.C.-dimension of $\mathcal{S}_{h_0}$ needs not to be h_0, it can be higher, because elements of different strata can build an antichain of size bigger than h_0. A further important point is that with this taming procedure we have introduced some asymmetry: In the case of the full closure system $\mathcal{S}$ of upsets it played no role that we looked at upsets and not at downsets: If we would have dealt with downsets to model the *poor* persons instead of modeling the *non-poor* persons via upsets, we would still have got the same results. The reason for this is simply that the complements of upsets are downsets and vice versa. In contrast to this, the complement of special selected upsets generated by antichains of some subset $T(h_0)$ of V are not necessarily downsets generated by the antichains of $T(h_0)$. Thus, for practical applications, one should analyze the results of both the tamed upset and the downset approach, which we will do in the example of application given in Section 6.1.

[32] If a taming with a global threshold δ appears more reasonable from a conceptual point of view in a concrete situation of application then a global taming may still be a better choice. However, in the example of application given in Section 6.1 we see no direct conceptual advantages of a global taming.

5.3.2 Taming formal contexts in the context of cognitive diagnosis models and knowledge space theory

For taming the closure system of all concept extents of a given formal context $\mathbb{K} = (G, M, I)$ with V.C.-dimension VC, in principal one can search for all shatterable objects sets of size VC (either step by step or in one whole step, see the remarks above) and exclude the objects of these sets from the context $\mathbb{K}$ to obtain a reduced context $\tilde{\mathbb{K}} = (\tilde{G}, M, I \cap \tilde{G} \times M)$ which then has a V.C.-dimension lower than VC (If this reduced V.C.-dimension is still too high, one can repeat the taming process until the resulting V.C.-dimension is low enough). For the actual data analysis one can then firstly take the closure system $\mathfrak{B}_2(\tilde{\mathbb{K}})$ of the intents of the reduced context $\tilde{\mathbb{K}}$ and secondly define the reduced closure system $\tilde{\mathcal{S}} := \{\{g \in G \mid \forall m \in B : gIm\} \mid B \in \mathfrak{B}_2(\tilde{\mathbb{K}})\}$ generated by all intents of the reduced context $\tilde{K}$ but w.r.t. the objects of the full original context $\mathbb{K}$. In Section 5.3.3 we will show how to do this in computational terms. Another possibility would be to not remove objects but attributes. In practical applications, often objects represent data points and the attributes represent the "multidimensional" values of the data points, so in classical situations one usually has much more objects than attributes. In these situations it appears more natural to remove objects, because if one would remove attributes, then one would remove these attributes for the whole big set of all objects. Compared to this, if one removes objects, then one removes only the specific concept intents generated by these objects (and also intents that are jointly generated by removed objects and non-removed objects). If one removes objects in the above described way, then one reduces the V.C.-dimension of the closure system under which the final analysis will be done. However, from a conceptual/descriptive/substance matter point of view, one does not know if one had removed sets that are actually interesting/important or that one did not remove uninteresting/unimportant sets. In some situations one can tame a context in a more guided manner:

One interesting example where one has some kind of substance matter guidance for taming is the case of **cognitive diagnosis models** (CDM), which are some kind of non-parametric item response models. Note that cognitive diagnosis models are very closely related to the theory of knowledge spaces ([Doignon and Falmagne, 2012], see [Heller et al., 2015]) which is itself closely related to formal concept analysis (see [Rusch and Wille, 1996]). In cognitive diagnosis models one has a set G of persons which respond to a set $M = \{1, \ldots, |M|\}$ of cognitive tasks, for example math tasks like fraction addition or fraction subtraction (for one well known fraction-subtraction data set see [Tatsuoka, 1984]). In contrast to more classical item response theory (IRT), in cognitive diagnosis modeling one is not mainly interested in measuring the abilities of persons and the difficulties of items, instead one is interested in the cognitive processes that generated the observed response patterns. Here, one demand is to give persons not only one or more numbers that measure their ability but to give a more qualitative feedback about which concrete skills the persons possess and which skills they do not possess. To do so, one develops (either theory driven or data driven or, in the best case, driven by a theory that

was rigorously empirically tested and persisted the tests) a so called Q matrix that specifies for every item, what kind of skills are in principle necessary to solve this item. Concretely, for a set of K relevant skills, a Q-matrix is a $|M| \times K$ matrix of zeros and ones where an entry $Q_{ij} = 1$ means that the skill j is needed to solve item i. In the simplest case one assumes that a person is expected to solve item i if she possesses all skills that are needed to solve item i, so a lack of one skill cannot be compensated by other skills. (This is the DINA model, cf. [Haertel, 1989, Junker and Sijtsma, 2001], but there are also other compensatory variants like the DINO model, cf. [Junker and Sijtsma, 2001].) Furthermore, one assumes the possibility of slipping an item one is principally prepared to solve and of luckily guessing the right answer of an item one is not prepared to solve. If for the moment we ignore the possibility of slipping and guessing, then the Q- matrix induces some structure of the idealized item response patterns that are possible if the probabilities of guessing and slipping are zero. For example if for solving one item i one needs all skills that one also needs for solving item i' plus some more, then response patterns of the form

$$(\ldots \underbrace{1}_{\text{i-th entry}} \ldots \underbrace{0}_{\text{i'-th entry}} \ldots)$$

are only possible due to a lucky guess of item i or a slipping of item i'. This fact can be expressed by saying that the formal implication $\{i\} \longrightarrow \{i'\}$ is valid in the closure system[33] $\mathcal{S}_Q := \mathfrak{B}_2(\{1,\ldots,K\},\{1,\ldots,|M|\},1-Q^T)$ of all possible idealized response patterns. To see that the closure system $\mathfrak{B}_2(\{1,\ldots,K\},\{1,\ldots,|M|\},1-Q^T)$ of the intents of the context $\mathbb{K}_Q := (\{1,\ldots,K\},\{1,\ldots,|M|\},1-Q^T)$ is exactly the space of all possible idealized response patterns, note that the intents are generated as $\{A' \mid A \subseteq \{1,\ldots,K\}\}$ where a set A can be understood as the set of skills an imaginary person does **not** posses. Then A' is the set of all items i with $\forall j \in A : (1-Q^T)_{ij} = 1$, i.e. the set of all items i where all skills the person does not possess are actually not needed to solve the item i. Thus, the intent A' is in fact the set of all items a person not possessing exactly all skills of A would actually be able to solve and all intents are exactly all observable idealized response patterns. A valid formal implication $Y \longrightarrow Z$ of $\mathbb{K}_Q$ could be interpreted in this situation as "All skills that are not necessary for solving any item from Y are also not necessary for solving items from Z" or alternatively as "every imaginary person who possess all skills for solving all items from Y also possesses all necessary skills for solving all items from Z".

Now, one can incorporate some or all valid implications of the idealized response pattern space $\mathcal{S}_Q$ to reduce the original closure system by looking only at concept intents of the original context $\mathbb{K} = (G,M,I)$ (where $gIm \iff$ person g has solved item m) that respect all or some of the valid implications of the formal context $\mathbb{K}_Q = (\{1,\ldots,K\},\{1,\ldots,|M|\},1-Q^T)$ representing the idealized response pattern space. If one enforces that all valid implications of the idealized response pattern space should also valid in the tamed closure system for the final analysis, then the V.C.-dimension would

[33] By abuse of notation, we identify the matrix $1-Q^T$ with the relation $\{(i,j) \mid (1-Q^T)_{i,j} = 1\}$.

decrease, but maybe unnecessarily too much. To enforce only a subset of implications one has to reasonably decide, which implications to include and which implications to not include. This can be made based on theoretical substance matter considerations about which implications are expected to be more clearly valid from a cognition theoretic perspective and which implications are maybe more questionable because they are due to a less rigorous but a more schematic specification of involved skills. To do so, one can substantially make use of the technique of **attribute exploration** (see [Ganter and Wille, 2012, p.85]) known from formal concept analysis: Given the formal context $\mathbb{K}_Q$ an algorithm like the next closure algorithm (see [Ganter and Wille, 2012, p.66-68]) can compute all formal concepts and also the so called stem base (see [Ganter and Wille, 2012, p.83]) of all valid implications of this context. In attribute exploration, at every step of the computation of a new implication, the user is asked in an interactive way, if the currently computed implication is actually true. Then the user can say that the implication is actually true or provide the algorithm with an object with specific attributes that are actually contradicting the formal implication. Then the algorithm would include this counterexample into the context and proceed, but not by computing all implications from the modified context anew, but by knowing that all implications computed before the counterexample was given are still valid in the modified context.

Another possibility of selecting implications to include for taming is to do it data driven. One can look for example at all valid implications $Y \longrightarrow Z$ of $\mathcal{S}_Q$ that are respected by at least a certain proportion C of objects from the original context $\mathbb{K}$ in the sense that at least a proportion C of persons, who solved all items of Y did also solve all items from Z. Formally, this can be described as enforcing all rules $Y \longrightarrow Z$ that have a so called confidence[34] $\mathrm{conf}(Y \longrightarrow Z)$ of at least C, where $\mathrm{conf}(Y \longrightarrow Z) := \frac{\mathrm{supp}(Y \cup Z)}{\mathrm{supp}(Y)}$ and $\mathrm{supp}(A) := |A'|$ and the operator $'$ is meant w.r.t. the original context $\mathbb{K}$. Here, the issue arises that if for example the rules $Y \longrightarrow Z_1$ and $Y \longrightarrow Z_2$ have a confidence above the threshold C then they would be included and furthermore the rule $Y \longrightarrow Z_1 \cup Z_2$ would implicitly be also valid in the tamed closure system, but this rule does not necessarily have a confidence of C. One can deal with this issue in different ways. One way of taming would be to enforce a set $\mathfrak{I}$ of implications that is deductively closed (this means that if an implication follows from some implications from $\mathfrak{I}$ then it should be already in the set $\mathfrak{I}$) and that only consists of implications with confidence above C and that is furthermore maximal w.r.t. these properties. Such maximal sets are generally not unique. Furthermore, if one has the idea that response patterns that violate implications of the idealized response pattern space are due to a random guessing or slipping, then if the slipping/guessing for different items is independent, for valid idealized implications $Y \longrightarrow Z_1$ and $Y \longrightarrow Z_2$ with confidence C_1 and C_2 one would expect a confidence of the implication $Y \longrightarrow Z_1 \cup Z_2$ that is generally lower than $\min\{C_1, C_2\}$. Thus, choosing the same threshold for implications with differently sized consequents seems to be not natural. Another way to proceed is to

[34] This term is used in the field of association rule mining, cf., e.g., [Agrawal et al., 1993, Piatetsky-Shapiro, 1991] which is also related to formal concept analysis, cf., e.g., [Lakhal and Stumme, 2005].

simply take the set of all implications with support above a threshold C and accept that with this set also implications from its deductive closure that may have a confidence smaller than C are also implicitly included for taming the closure system. If one takes this route, one is faced with computing all implications with confidence above C. For an implication $Y \longrightarrow \{z_1, z_2\}$ with confidence above C it is necessary that also the implications $Y \longrightarrow \{z_1\}$ and $Y \longrightarrow \{z_2\}$ have confidence above C, so it suffices to look only at implications with a singleton consequence. (The implication $Y \longrightarrow \{z_1, z_2\}$ is included for taming if and only if both $Y \longrightarrow \{z_1\}$ and $Y \longrightarrow \{z_2\}$ have confidence above C because this is necessary and if both $Y \longrightarrow \{z_1\}$ and $Y \longrightarrow \{z_2\}$ have confidence above C they are both included and thus also $Y \longrightarrow \{z_1, z_2\}$ has to be included since it follows from the included implications $Y \longrightarrow \{z_1\}$ and $Y \longrightarrow \{z_2\}$.) Instead of computing all implications with a singleton consequence one can also compute in a first step only that implications that have a minimal antecedent. Then one could exclude such implications $Y \longrightarrow \{z\}$ with confidence lower than C and recompute the generic base. To do so, one can directly work with the formal context $\mathbb{K}_Q := (\{1, \ldots, K\}, \{1, \ldots, |M|\}, 1 - Q^T)$. One can compute the generic basis and split every implication $Y \longrightarrow \{z_1, \ldots, z_l\}$ into implications $Y \longrightarrow \{z_1\}, \ldots, Y \longrightarrow \{z_l\}$ with singleton consequents. Then, for every such implication $Y \longrightarrow \{z\}$ with confidence lower than C one can exclude it by adding a the counterexample $Y''\backslash\{z\}$ as a further item pattern to the context $\mathbb{K}_Q$. (Here the operation $''$ is meant w.r.t. the context $\mathbb{K}_Q$.) Then one can compute again and again the generic base of the enlarged context until no rule has confidence lower than C anymore. A computationally more elegant way would be to not to recompute the whole rule base of the enlarged context anew. In the spirit of attribute exploration (see [Ganter and Wille, 2012, p.85]), one can smartly exclude implications with confidence lower than C directly during the generation of the rules. However, since one does not work with the generic base, but with the stem base, the result would then be different and would furthermore dependent one the concrete order in which the computed implications were presented to the user.

5.3.3 How to compute the test statistics for the tamed closures systems

In this section we shortly indicate, how one can compute the test statistic for a tamed closure system. We start with the example of the closure system of upsets. In Section 5.3.1 we came up with the tamed family of sets $\mathcal{F}_{h_0} = \{\uparrow B \mid B \subseteq T(h_0)\}$ that generates the closure system $\mathcal{S}_{h_0} = \mathrm{cl}(\mathcal{F}_{h_0})$. Of course, it would be intractable to explicitly compute the closure system $\mathcal{S}_{h_0}$ generated by $\mathcal{F}_{h_0}$ because it is simply too big. Fortunately, the explicit computation of $\mathcal{S}_{h_0}$ is not needed: The closure system $\mathcal{S}_{h_0}$ simply consists of all possible intersections of sets of the generating family of sets $\mathcal{F}_{h_0}$. For ease of presentation, assume that $(V, \leq)$ itself is already a complete lattice (otherwise, simply take its Dedekind-MacNeille completion, cf., e.g., [Ganter and Wille, 2012, p.48]). The closure system $\mathcal{S}_{h_0}$ is simply the set of all possible intersections of sets of the family $\mathcal{F}_{h_0}$. For a finite[35] family $(\uparrow A_i)_{i \in \{1,\ldots,n\}}$ of upsets from $\mathcal{F}_{h_0}$, the intersection $\uparrow A_1 \cap \ldots \cap \uparrow A_n$

[35] The finiteness is actually not needed, it only makes the presentation more simple, here.

can be written as $\uparrow A_1 \cap \ldots \cap \uparrow A_n = \bigcup\{\uparrow a_1 \cap \ldots \cap \uparrow a_n \mid \forall i \in \{1,\ldots,n\} : a_i \in A_i\} = \bigcup\{\uparrow \bigvee\{a_1, \ldots a_n\} \mid \forall i \in \{1,\ldots,n\} : a_i \in A_i\}$. Thus, $\mathcal{S}_{h_0}$ can be written as $\mathcal{S}_{h_0} = \{\uparrow B \mid B \subseteq \bar{T}(h_0)\}$ where $\bar{T}(h_0) = \{\bigvee A \mid A \subseteq T(h_0)\}$. Since $\bigcup_{i\in I} \uparrow B_i = \uparrow \bigvee_{i\in I} B_i$ for arbitrary families $(B_i)_{i\in I}$, the closure system $\mathcal{S}_{h_0}$ is closed under arbitrary unions and thus, because of Birkhoffs theorem, the valid implications of $\mathcal{S}_{h_0}$ are simple implications. Thus, we can firstly calculate all simple implications or a basis thereof and implement them in a linear program: For example one can compute for every $x \in V$ the set $\downarrow x \cap T(h_0)$ of all elements of $T(h_0)$ that are below x. Then, one can take from the set M of all upper bounds of $\downarrow x \cap T(h_0)$ the minimal elements $\min M$. Finally, for every $y \in \min M$ one simply has to implement the associated implication $\{x\} \longrightarrow \{y\}$ as an inequality constrain in the linear program.

For the case of non-guided taming of a closure system that is given by a generating formal context, remember that the taming was simply done by removing objects from the context (but only for the generation of the closure system of the intents, and not for the whole analysis). Let I denote the set of indices of the objects that were excluded for the generation of the closure system. To compute the statistic for the tamed context, one only has to modify the program (24) to the following program:

$$\begin{aligned} \langle (w_1^{ext}, \ldots, w_m^{ext}, w_1^{int}, \ldots, w_n^{int}), (z_1, \ldots, z_m, z_{m+1}, \ldots, z_{m+n}) \rangle &\longrightarrow \max \\ w.r.t. & \\ \forall (i,j) \text{ s.t. } A_{ij} = 0 : z_i &\leq 1 - z_{j+m} \\ \forall i \in \{1,\ldots,m\} : \sum_{k:A_{ik}=0} z_{k+m} &\geq 1 - z_i \\ \forall j \in \{1,\ldots,n\} : \sum_{k\notin I : A_{kj}=0} z_k &\geq 1 - z_{j+m} \end{aligned} \tag{39}$$

Here, the only difference is that in the last set of inequalities, one does not sum over every object index k but only over that indices, that were not excluded for the generation of the closure system. To see the validity of this modification, simply note that the three verbalizations directly above the linear program (24) are still exactly characterizing the situation with the only modification of point 3, which has to be modified to

"Dually, if attribute m_j does not belong to the intent, then there exists at least one object g_k that was not excluded for the generation of the closure system of intents, and that belongs to the extent, but does not have attribute m_j."

For the guided taming the computation of the test statistic is straightforward. Since the formal implications one additionally imposes are computed explicitly, one can modify the binary program described in Section 4.2 by additionally implementing the further imposed implications as inequality constraints like described in Section 4.1.

6 Examples of application

In this section we apply the developed methods to different data sets. The applications should on the one hand be not taken at face value as serious substance matter applications. On the other hand, they should also be not misunderstood as pure toy examples. The aim of the following examples of application is to show that the developed methods are in fact applicable to "real-world" data sets and that these methods are in principle very flexible and can also deal with different kinds of data deficiency. The big part that is missing to make the examples serious substance matter studies is the fact that at much stages of the analysis, some substance matter considerations have to be made or could maybe be made to make the analysis more decisive. However, since the authors are clearly no experts in the substance matter fields the applications are related to, they would like to refrain from making such substance matter decisions, if possible, or to make the actually needed substance matter decisions only for purposes of illustration. In the following examples, especially in our main example of Section 6.1, we analyze the data sets by a more generic way of proceeding and, if appropriate, shortly indicate, at which steps and in which way one could make a more refined data analysis, that of course would be dependent on some substance matter decisions.

6.1 Upsets: Relational inequality analysis

We start with our main example of multivariate inequality analysis using data from the German General Social Survey (ALLBUS) of the year 2014 (GESIS - Leibniz - Institut fur Sozialwissenschaften [2015]). In this survey, altogether 3471 persons participated. Here, we analyze systematic multivariate differences between the group of male and female participants w.r.t. the variables *Income*, *Education* and *Health*. The question about *Health* was asked in a split ballot design to test for a possible impact of different response scales on the result. The participants were asked both in split A and split B about how they would describe their health status in general. The participants of split A got the 5 different answer categories "Sehr gut" (very good), "Gut" (good), "Zufriedenstellend" (satisfactory) , "Weniger gut" (suboptimal) and "Schlecht" (bad) whereas the participants of split B got the additional category "Ausgezeichnet" (excellent). (The english categories in brackets are our own english translation.) For reasons of simplicity, we used here only the participants of split B and did a complete case analysis.

Of course, one could also use both splits for the analysis: If one has some reason to assume that both response scales adequately operationalize the same construct, one can do a joint analysis of both splits by matching the two scales to each other based on their respective empirical distribution functions. This is actually possible because the splitting was random and thus the measured construct has the same distribution in every split.[36]

[36]Note that due to measurement error, which can be different within the two splits, the actual measurements can differ in their distribution. But if the measurement errors are independent of the measured construct and from each other, this will only produce some "smearing" of the measurements, which can

For the joint analysis of the three variables *Income*, *Education* and *Health*, the complete case analysis consisted of altogether 1515 participants (706 female and 809 male) corresponding to a non-response rate of 12.2%. The variable *Income* contributed most to the non-response-rate (the non-response rate for *Income* was 11.8%.) Here, income was asked for in a two step procedure: First with an open question and then, for participants who refused to answer the open question, a categorized question with 23 answer-categories ranging from "no income" to "more than 7500 Euro" was added. This two-step procedure was done to reduce the non-response rate. Here, for simplicity we use the combined answers to the open and the list query, where for participants who answered only the list query simply the mid-points of the interval representing the categorized answer were used as a surrogate for the true income[37]. Note that for our analysis we only need the ordinal structure of the variable *Income* and furthermore we can actually deal also with a partially ordered structure of the dimension income. Thus, here one can also use more cautious approaches where one says for example that an income that is actually only categorically observed as $[a, b)$ is only lower than or equal to another observed income (no matter if precisely observed as $[c, c]$ or imprecisely observed as $[c, d)$) of $[c, d)$ iff $b \leq c$. Another possibility would be to say that categorically observed incomes $[a, b)$ are comparable to itself (i.e. $[a, b) \leq [a, b)$), but not to a precisely observed value $c \in [a, b)$). The stochastic dominance approach is thus very flexible to deal with certain kinds of non-response/interval-valued observations. Here, we do simply work with the combined values where interval-valued observed incomes are replaced by the corresponding interval mid-points.

The variable *Education* is the classification of the level of education according to the International Standard Classification of Education (ISCED) 2011 (see [UNESCO Institute for Statistics (UIS), 2012]) implemented for Germany. On the highest stage, this classification differentiates between 9 different main levels of education:

Level 0:	Less than primary education	Level 5:	Short-cycle tertiary education
Level 1:	Primary education	Level 6:	Bachelor's or equivalent level
Level 2:	Lower secondary education	Level 7:	Master's or equivalent level
Level 3:	Upper secondary education	Level 8:	Doctoral or equivalent level
Level 4:	Post-secondary non-tertiary education		

We treat here the variable *Education* as of totally ordered scale of measurement. In the sample, only the levels from 1 to 8 were observed. Note that also for this dimension the methodology of stochastic dominance would be able to deal with an only partially ordered scale: The ISCED 2011 could also be implementation in a more cautious way: For example, instead of only comparing the highest educational achievements, one could

lead to cases where stochastic dominance w.r.t. the underlying construct is actually present, but it is not present anymore for the measurements. A transition of non-stochastic dominance w.r.t. the construct into stochastic dominance w.r.t. the measurements cannot happen.

[37] For the answer category "below 200 Euro" a value of 150 Euro and for the category "more than 7500 Euro " a value of 8750 Euro was assigned.

alternatively look at the whole educational paths and and say that a person A is more "poor" than another person B w.r.t. the dimension *Education* only if both persons followed the same educational path but person A stopped earlier with a lower highest educational achievement than person B. This partial ordering of the dimension *Education* would lead to a less decisive analysis, but it has the potential to reveal, how much a more classical analysis would dependent on the choice of a totally ordered scale for the dimension*education*.

We begin with a marginal analysis of all 3 variables. Figure 3 shows the lower cumulative distribution function for every variable for both the male and the female group. One can see that the female group is almost dominated by the male group for the variables *Income*, *Education* and *Health*. With regard to *Income*, the extent of dominance is the highest: 66.4% of the women earn not more than 1300 Euro, but only 31.9% of the men earn not more than 1300 Euro, which is a difference of 34.5 percentage points. Only for the very high income of 12000 Euro there is a small deviation from dominance in the sense that 99.9% of the men earn not more than 12000 Euro, where this is the case for only 99.8% of the women. For the variable *Health* there is only deviation from dominance w.r.t. the percentage of women reporting a health-status *bad*: Only 2.2% of the women report a health status *bad*, which is about 0.7 percentage points lower than the amount of 2.9% for the men. The variable *Education* shows strict dominance.

Now, let us come to the joint analysis. For the statistics

$$\begin{aligned} D^+ &= \max_{M \in \mathcal{U}((V,\leq))} \langle w^x - w^y, m \rangle \\ D^- &= \min_{M \in \mathcal{U}((V,\leq))} \langle w^x - w^y, m \rangle, \end{aligned}$$

where X describes the subpopulation of male, and Y describes the subpopulation of female persons, we obtain

$$\begin{aligned} D^+ &\approx 36.48\% \\ D^- &\approx -1.21\%, \end{aligned}$$

which indicates an almost strict dominance for the joint distribution of the variables *Income*, *Education* and *Health*, where the small deviation from dominance is with $D^- \approx -1.21\%$ not much higher than the largest deviation of -0.7% for the variable *Health* in the marginal analysis. The maximal value of 36.48% is about 2 percentage points higher than for the largest maximal value of 34.5% for the variable *Income* in the marginal analysis. Beyond the purely quantitative analysis one can also look, at which upsets the maximum and the minimum of the test statistic is attained. The maximum of the statistic is attained at an upset U generated by the antichain A (via $U = \uparrow A$) containing 9 elements depicted in Table 3.

The minimal test statistic is attained at an upset generated by an antichain of size 4 described in Table 4. Based on a resampling scheme with 10000 replications, the test sta-

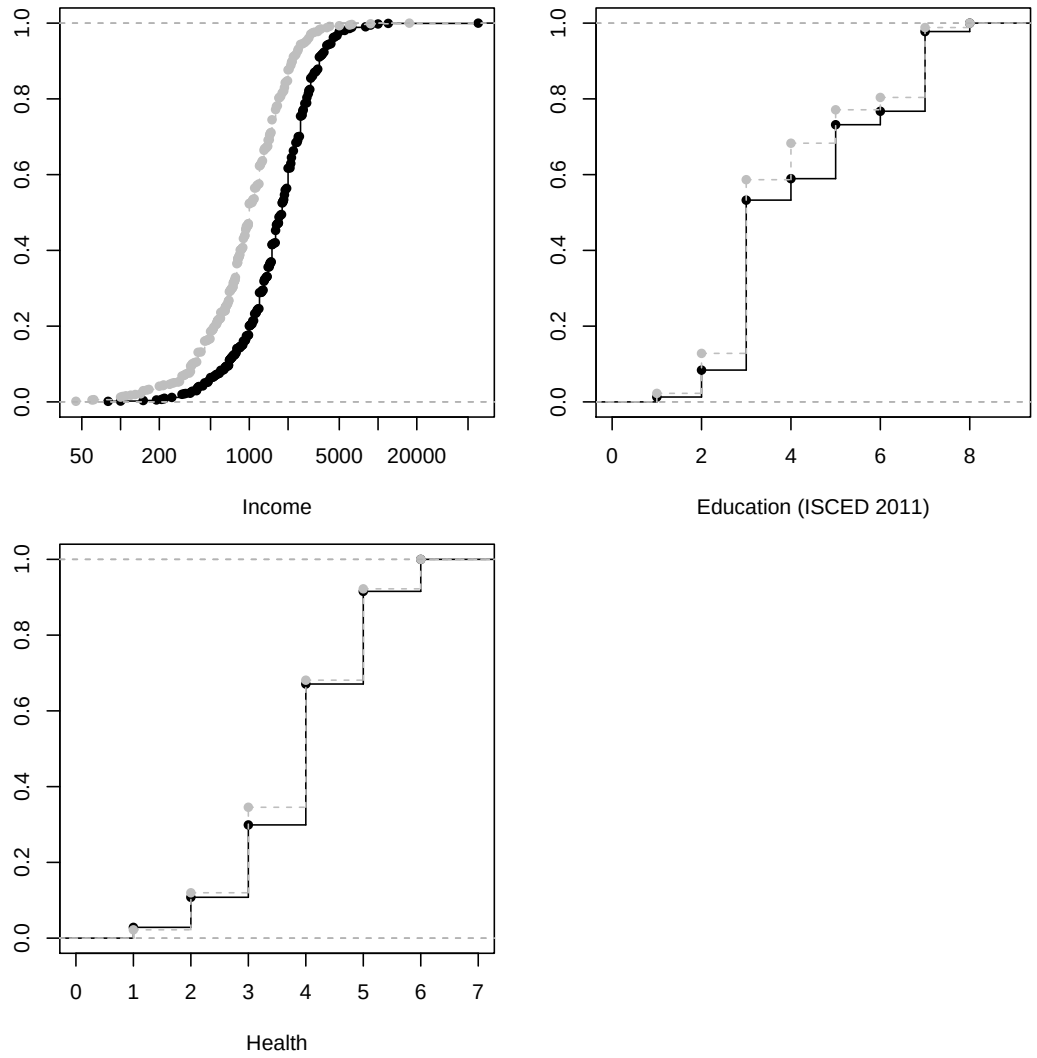

Figure 3: Empirical cumulative distribution function for all 3 considered variables for the male group (black) and the female group (grey).

tistic D^+ appears as highly significantly above 0 whereas D^- is really only non-significantly different from zero: The maximal observed value of D^+ in the resample is 17.48% and the minimal value of D^- observed in the resample is -1.63%, which is very close to -1.21% for the actually analyzed data set, actually, the value of -1.21% is closer to zero for the actual data set than the closest value of the resample. The poset generated by the actually observed data and the coordinate-wise ordering has a Vapnik-Chervonenkis-dimension (width) of 33. For such a V.C.-dimension and an n around[38] 700, the V.C.-inequality

[38] The data set included 706 female and 809 male participants. Note that actually the sampling weights for east and west have to be also taken into account, here, however, we only want to get a rough idea about how sharp the V.C.-inequality is in our situation.

	Income (Euro)	Education (ISCED 2011)	Health (self-reported)	difference	above
1	400 (0.93)	Upper secondary education (0.9)	excellent (0.08)	0.02	0.06
2	650 (0.84)	Lower secondary education (0.99)	excellent (0.08)	0.02	0.06
3	1080 (0.64)	Master's or equivalent level (0.22)	very good (0.32)	0.02	0.07
4	1100 (0.64)	Master's or equivalent level (0.22)	good (0.68)	0.06	0.14
5	1260 (0.55)	Master's or equivalent level (0.22)	satisfactory (0.89)	0.08	0.17
6	1300 (0.55)	Upper secondary education (0.9)	satisfactory (0.89)	0.3	0.49
7	1400 (0.51)	Primary education (1)	good (0.68)	0.25	0.37
8	1400 (0.51)	Upper secondary education (0.9)	bad (1)	0.33	0.49
9	1450 (0.48)	Lower secondary education (0.99)	satisfactory (0.89)	0.32	0.45

Table 3: The antichain $A = \{A_1, \ldots, A_9\}$ that generates that upset $U = \uparrow A$ where the maximum of the test statistic is attained. In brackets the marginal upper quantiles that correspond to the values are given, e.g. the 0.93 behind the 400 in the first row of the first column means that ca. 93% of the persons in the population earn at least 400 Euro. The column *difference* displays for every row i the difference between the proportion of male and the proportion of female persons that are above element A_i. The column *above* shows the proportion of all persons that are above A_i.

	Income (Euro)	Education (ISCED 2011)	Health (self-reported)	difference	above
1	100 (1)	Master's or equivalent level (0.22)	excellent (0.08)	-0.0019	0.02
2	130 (0.99)	Upper secondary education (0.9)	suboptimal (0.97)	0.0359	0.87
3	600 (0.86)	Lower secondary education (0.99)	very good (0.32)	0.0759	0.27
4	2900 (0.12)	Master's or equivalent level (0.22)	bad (1)	0.0797	0.07

Table 4: The antichain $A = \{A_1, \ldots, A_4\}$ that generates that upset $U = \uparrow A$ where the minimum of the test statistic is attained. In brackets the marginal upper quantiles that correspond to the values are given, e.g. the 1 behind the 100 in the first row of the first column means that ca. 100% of the persons in the population earn at least 100 Euro. The column *difference* displays for every row i the difference between the proportion of male and the proportion of female persons that are above element A_i. The column *above* shows the proportion of all persons that are above A_i.

(26) is too loose. For a value of the test statistic of about 36% one would have to have chosen a V.C.-dimension of about 8 to make the conservative V.C.-inequality leading to a significant result. Since we were able to compute a large enough resample, we actually do not need to rely on the V.C.-inequality. However, for the purpose of illustration, we can tame the closure system of upsets to get an insight into how this affects the behavior of the test statistic for the actually observed data and the distribution of the test statistic under H_0. Figure 4 shows a the value of the test statistic D^+ for the actually observed data, as well as the distribution[39] of D^+ under H_0 for different V.C.-dimensions ranging from 4 to 39. Note that the original V.C.-dimension was 33, which is maybe surprising, but the V.C.-dimension of 39 for the biggest tamed closure system is due to the fact that by

[39]Here, we computed a resample of size 1000 to get a rough insight into the distribution of D^+ under H_0.

taming the closure system one gets in a first step only a family of sets that is generally no closure system and one has to enlarge this family in a second step to be a closure system to make the analysis computationally feasible. One can see that, as expected, with increasing V.C.-dimension, both the value of the test statistic for the actually observed data, as well as the expectation of the test statistic under H_0 increases. The standard deviation of the test statistic has also an increasing trend for increasing V.C.-dimensions. If one standardizes the test statistic D^+ by subtracting its mean and dividing the result by its standard deviation, one sees that the shape of the distribution of D^+ is approximately independent of the V.C.-dimension. The fact that the shape of the test statistic is approximately independent of the V.C.-dimension could possibly be used to get rules of thumb for situations where the computation of large resamples is computationally intractable. However, the approximate independence of the shape of the distribution of D^+ from the V.C.-dimension may be only present in our special situation and thus may be misleading for getting a rule of thumb for the general case.

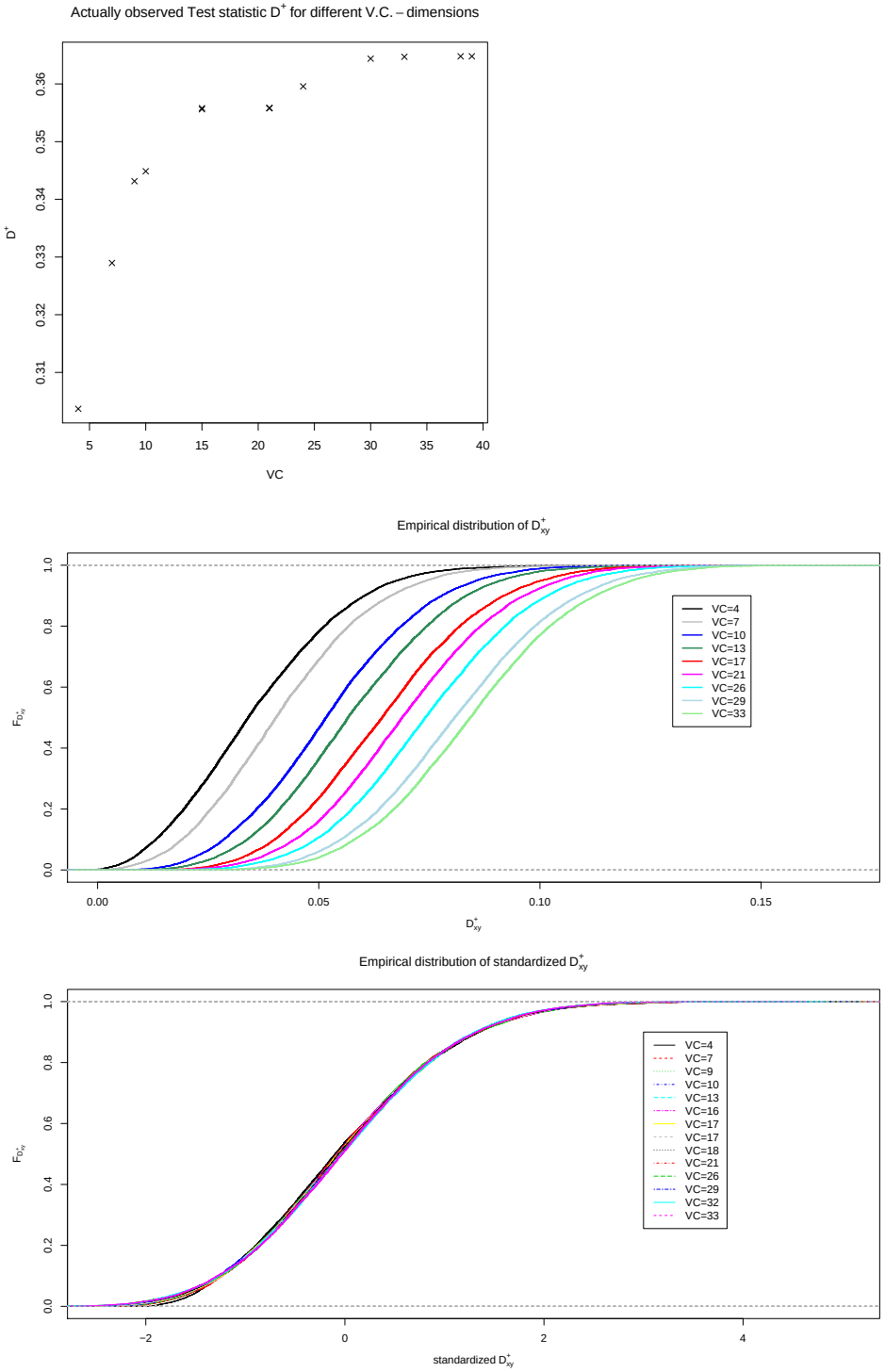

Figure 4: The value of the test statistic D^+ for the actually observed data, as well as the distributions of the test statistic D^+ and the standardized test statistic $\frac{D^+ - \overline{D^+}}{sd(D^+)}$ under H_0 for different V.C.-dimensions. One can see that the shape of the distribution D^+ is nearly independent of the V.C-dimension.

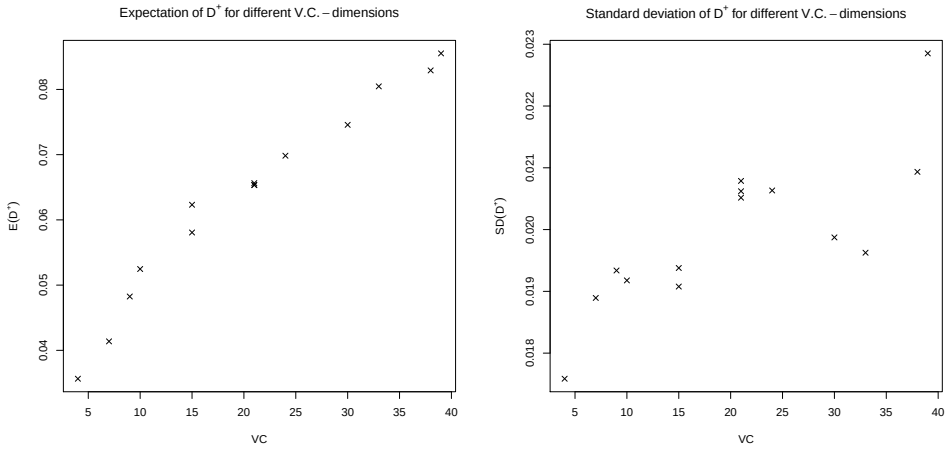

Figure 5: The expectation and the standard deviation of D^+ under H_0 for different V.C.-dimensions.

Now, we would like to illustrate a little bit, how the taming behaves w.r.t. conceptual terms. Therefore, we analyze for a tamed closure system of V.C.-dimension 7 the tamed upsets and downsets, where the maximum D^+ and the minimum D^- is attained. For the upset-approach, the maximal statistic for the tamed closure system is attained at an antichain of size 7 summarized in Table 5.

	Income (Euro)	Education (ISCED 2011)	Health (self-reported)	difference	above
1	1020 (0.65)	Short-cycle tertiary education (0.37)	excellent (0.08)	0.01	0.02
2	1050 (0.65)	Short-cycle tertiary education (0.37)	very good (0.32)	0.04	0.12
3	1050 (0.65)	Master's or equivalent level (0.22)	good (0.68)	0.06	0.14
4	1063 (0.65)	Short-cycle tertiary education (0.37)	good (0.68)	0.11	0.23
5	1200 (0.6)	Upper secondary education (0.9)	good (0.68)	0.23	0.42
6	1248 (0.56)	Upper secondary education (0.9)	satisfactory (0.89)	0.3	0.5
7	1300 (0.55)	Upper secondary education (0.9)	suboptimal (0.97)	0.32	0.52

Table 5: The antichain $A = \{A_1, ..., A_7\}$ that generates that upset $U =\uparrow A$ where the maximum of the test statistic is attained for the tamed closure system with a V.C. dimension of 7.

One can see that the maximal difference between the transformed Z-values in brackets is $0.65 - 0.08 = 0.57$ attained for the first element A_1, where the Z-value of 0.65 for an income of 1020 Euro is the largest, and a Z-value of 0.08 for a health status *excellent* is the smallest value. Compared to this, in the non-tamed case, the skewest element of the antichain generating the upset where the maximal value of the test statistic is attained is the element A_2 with a maximal Z-value of 0.99 for an education status *Lower secondary education* and a minimal Z-value of 0.08 for a health status *excellent*. The difference $0.99 - 0.08 = 0.91$ is clearly greater than for the tamed situation showing that we actually managed to reduce the skewness of elements generating the closure system for the tamed analysis.

The value of the tamed test statistic is with 32.90% not much smaller than the initial value of 36.48%, still significantly different from zero. The minimal value is -0.045% attained at the antichain consisting of only one element depicted in Table 6

	Income (Euro)	Education (ISCED 2011)	Health (self-reported)	difference	above
1	3500 (0.08)	Master's or equivalent level (0.22)	excellent (0.08)	-5e-04	0.003

Table 6: The antichain $A = \{A_1\}$ that generates that upset $U =\uparrow A$ where the minimum of the test statistic is attained for the tamed closure system with a V.C. dimension of 7.

The minimal value is not significantly different from zero. Table 7 and Table 8 finally show the results one would obtain if one would do the tamed analysis by looking at downsets instead of upsets:

Obviously, the role of the maximal and the minimal value of the test statistic will interchange: The maximal value of the test statistic of 0.15% means that the difference between the proportion of the *poor* male and the *poor* female persons is maximally 0.15% attained if one concretizes the term *poor* with the downset $D =\downarrow A$ generated by the antichain given in Table 7. The minimal value of the test statistic is -28.82% attained for the downset generated by the antichain given in Table 8. Note that for the downset-analysis we used for the construction of the Z-values not the complementary distribution function, but the usual distribution function, because this fits better to the notion of a downset. This only has an impact on the interpretation of the numbers given in brackets. For example the 0.06 beyond the income value of 360 Euro in Table 7 means now, that 6% of the population have income *below* 360 Euro. Additionally, the last column, denoted *below*, now gives the proportion of persons *below* the corresponding element of the antichain and the column *difference* gives the difference of the proportions below the corresponding element.

	Income (Euro)	Education (ISCED 2011)	Health (self-reported)	difference	below
1	360 (0.06)	Primary education (0.01)	bad (0.03)	0.0015	0.0008

Table 7: The antichain $A = \{A_1\}$ that generates that downset $D =\downarrow A$ where the maximum 0.0015 of the test statistic is attained for the tamed closure system with a V.C. dimension of 7.

6.2 Concept extents: Gender differences and differential item functioning in an item response dataset

In this section, we shortly analyze an IRT-dataset w.r.t. gender differences and **Differential item functioning** (DIF, [Osterlind and Everson, 2009]). The data set is a subsample from the general knowledge quiz *Studentenpisa* conducted online by the German weekly

	Income (Euro)	Education (ISCED 2011)	Health (self-reported)	difference	below
1	1400 (0.51)	Bachelor's or equivalent level (0.78)	good (0.68)	-0.21	0.33
2	1450 (0.52)	Short-cycle tertiary education (0.75)	very good (0.92)	-0.28	0.41
3	1460 (0.52)	Post secondary non-tertiary education (0.63)	good (0.68)	-0.20	0.3
4	1474 (0.53)	Upper secondary education (0.55)	good (0.68)	-0.16	0.28

Table 8: The antichain $A = \{A_1, ..., A_4\}$ that generates that downset $D = \downarrow A$ where the minimum -0.288 of the test statistic is attained for the tamed closure system with a V.C. dimension of 7.

news magazine SPIEGEL ([SPIEGEL Online, 2009], see also Trepte and Verbeet [2010] for a broad analysis and discussion of the original data set.) The data contain the answers of 1075 university students from Bavaria to 45 multiple choice items concerning the 5 different topics *politics*, *history*, *economy*, *culture* and *natural sciences*. For every topic, 9 questions were posed, for example question 1 of the politics topic was: "Who determines the rules of action in German politics according to the constitution?". The data set was analyzed in a number of papers, for example in Strobl et al. [2015], Tutz and Schauberger [2015], Tutz and Berger [2016], mostly from an IRT point of view. All mentioned papers identified systematic differences between the subgroups of male and female students in the sense of the presence of differential item functioning. Differential item functioning is present if the distribution of the item response patterns in two subgroups with identical latent abilities are different. Here, one cannot assume that the subgroups of male and female students that actually participated in the online quiz have the same latent abilities, because for example self selection processes can be present. To analyze the presence of differential item functioning one has to firstly somehow match persons of the two subgroups with similar abilities. One classical non-parametric procedure is the test of Mantel Haenszel (see [Holland et al., 1988].), where one takes the item scores (i.e., the number of solved items) as a matching criterion[40]. One stratifies the populations into parts with the same item score and then compares the subpopulations in every stratum. The final test statistic is then a χ^2-type statistic cumulating over all strata. The Mantel Haenszel procedure is an item-wise test, one tests for every item separately, if DIF is present for this item. For the construction of the matching score one usually does not take the whole set of items, instead one ignores items that showed DIF in a first preliminary analysis that was based on the whole set of items[41]. This process is called purification and there are different variants of purification, see, e.g., [Osterlind and Everson, 2009, p.16]. We can use the linear programming approach on formal contexts to develop a joint DIF test based on the item scores as a matching criterion. Firstly, we have to care for the different distributions of the abilities in the different subgroups. Here, we do not make a conditional analysis since conditioning would make all classes with the same item score relatively small such that a 45-dimensional multivariate

[40]Note that this will only work if the score values are a sufficient statistic for the abilities, which is for example the case for the Rasch model. For a discussion of deviations from this assumption in the context of the classical Mantel Haenszel procedure, see, e.g., [Zwick, 1990]

[41]The actually tested item should always be included for matching to make the Mantel Haenszel procedure valid under the null hypothesis of no differential item functioning, see [Holland et al., 1988, p.16].

analysis in every stratum would expectedly have very low power. Instead, we re-weight both subgroups such that the ability distributions in the male and the female group are approximately the same and then we analyze the joint distribution of item patterns and abilities (measured via the item scores). Concretely, we do the following:

1. Let $\mathbb{K}_0 = (G, M, I)$ be the formal context where $G = \{g_1, \ldots, g_{1075}\}$ is the set of persons, $M = \{m_1, \ldots, m_{45}\}$ is the set of items and gIm iff person g solved item m.

2. Separately for the male and the female group we estimate the density of the distribution of the item scores s, denoted with $\hat{f}_{male}$ and $\hat{f}_{female}$, respectively. The estimation is done here with a kernel density estimator.

3. Then we inversely re-weight the sample by giving a weight

$$W_i := \begin{cases} \hat{f}_{male}(s_i) & \text{if the } i\text{th person is female} \\ \hat{f}_{female}(s_i) & \text{if the } i\text{th person is male.} \end{cases}$$

   After this, the re-weighted distribution of the scores in the male and female group are approximately the same.

4. Then we analyze the joint distribution of response patterns and the score values in both subgroups. To do this, we use the flexibility of formal context analysis and simply conceptually scale the score values with an interordinal scale. Concretely, for every score value s we add an attribute "$\leq s$" and an attribute "$\geq s$" to the original context $\mathbb{K}_0$ with the interpretation person g has attribute "$\leq s$" if g has a score value lower than or equal to s and person g has attribute "$\geq s$" if g has a score value greater than or equal to s. Afterwards, we analyze the enlarged context $\mathbb{K}_1$ by looking at the closure system $\mathfrak{B}_1(\mathbb{K}_1)$ and computing

$$\max/\min_{M \in \mathfrak{B}_1(\mathbb{K}_1)} \langle (w^x - w^y) \cdot W, \mathbb{1}_M \rangle,$$

   where w^x are the original weights for the male and w^y are the weights for the female persons. Concretely, in the sample there were 658 male and 417 female persons, thus

$$w_i^x = \begin{cases} \frac{1}{658} & \text{if the ith person is male} \\ 0 & \text{if the ith person is female} \end{cases}$$

   and

$$w_i^y = \begin{cases} 0 & \text{if the ith person is male} \\ \frac{1}{417} & \text{if the ith person is female} \end{cases}.$$

5. In a last step we apply a purification procedure by basing the item scores for matching only on items that are not in the concept intents for which the maximal and minimal test statistic was obtained in a first run. We repeat the purification procedure until not further items are excluded.

Before showing the actual results, we firstly compute the test statistics D^+ and D^- for the context $\mathbb{K}_0$ without re-weighting the data. The context has a V.C.-dimension of 22 and has about 8.900.000.000 formal concepts (this is an estimate based on random sampling of arbitrary item sets and checking if they are a concept intent) and is thus very hardly describable explicitly and we will use the binary program described in Section 4.2 to compute the test statistics [42]. The maximal value of the test statistic is 0.335 attained at a formal concept containing the questions

F6: "Who is this? - (Picture of Horst Seehofer.)"

F26: "Which internet company took over the media group Time Warner? - AOL."

This means that the difference in the proportions of male and female persons who answered at least questions $F6$ and $F26$ correctly is the greatest observed difference between proportions of male and female persons that answered at least all items of some set of items correctly. Concretely, 53.6% of the male and 20.1% of the female persons answered these both questions rightly. The minimal value of the test statistic is -0.169 attained at a formal concept containing the questions

F40: "What is also termed Trisomy 21? - Down syndrome."

F43: "Which kind of bird is this? - Blackbird."

Here, 59.6% of the male and 76.5% of the female persons got both questions right. Both differences are significant, for a resample of size 1000 the value $\max\{D^+, -D^-\}$ had a range from 0.05 to 0.14 and a standard deviation of 0.014.). Figure 6 gives a rough idea about how a (non-guided) taming of the closure system by removing big shatterable sets of objects from the context affects the distribution of the test statistic D^+ under H_0. (Note, that this is only a very small simulation where we resampled only 100 times for every value of the V.C.-dimension.) The initial context has a V.C.-dimension of 22. One can see, that by reducing the V.C.-dimension, the mean and the standard deviation of the test statistic does not change very much for small reductions of the V.C.-dimension. Only a very strong taming to a V.C.-dimension below 8 seems to have an effect in reducing the mean of D^+ under H_0.

Now we come to the actual DIF-analysis: Figure 7 shows the distribution of the item scores for the male and female persons. The distributions are very different and thus we have to correct for this difference by re-weighting the data. However, generally, every attempt to account for such a kind of difference should be taken with some grain of salt, because initially we would like to account for differences in the abilities, but the abilities are only latent traits that cannot be observed and thus have to be estimated, in our situation

[42] To get a rough idea of computational complexity: The MIP solver `Gurobi` (see [Gu et al., 2012]) needed ca. 100 seconds to compute the statistic using one core on a 2.60 Ghz CPU (Intel(R) Xenon(R) CPU E5-2650 v2 @2.60 Ghz, 64GB RAM).

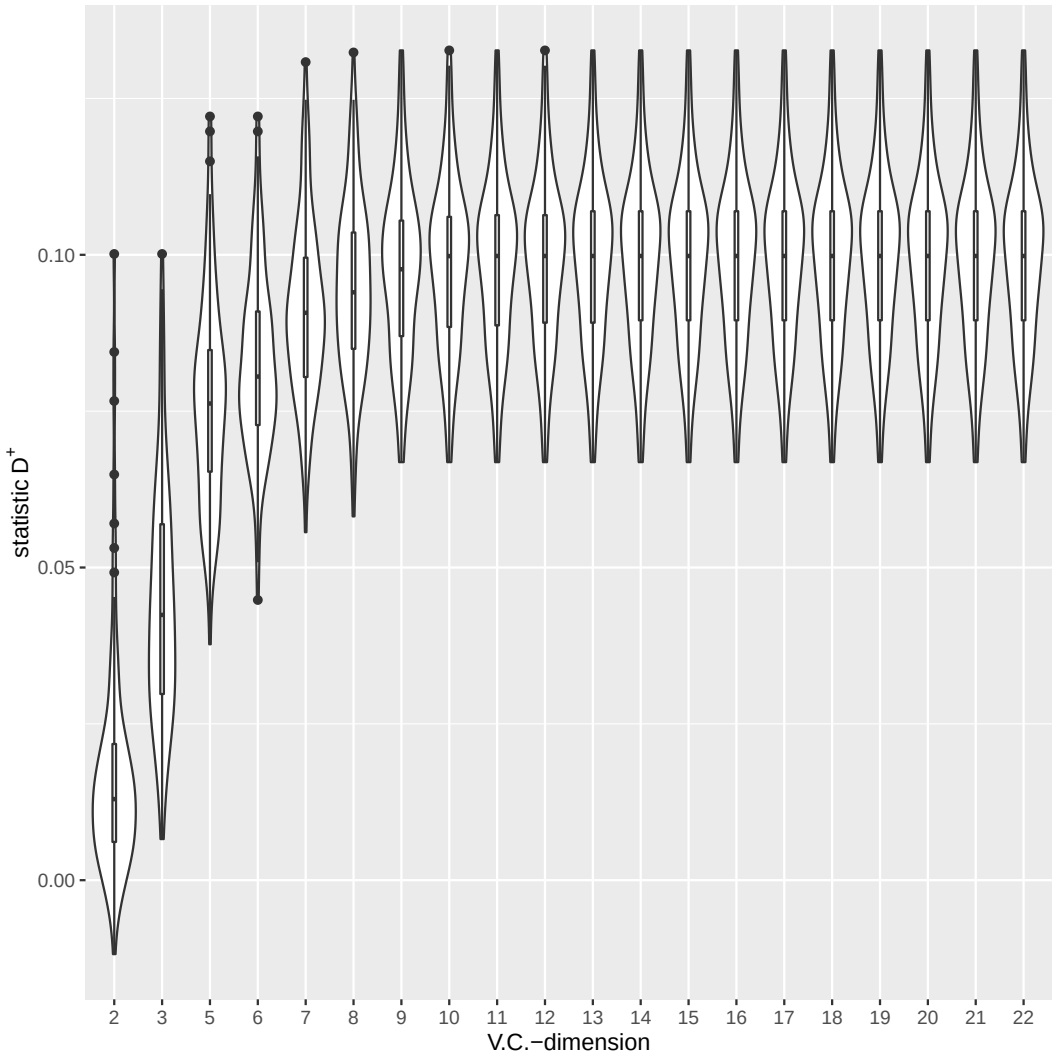

Figure 6: Distribution of the tamed statistic D^+ for different V.C.-dimensions.

from the item scores. In the unlucky case, an attempt for accounting for differences in abilities can make the analysis still more misleading if the items that suffer from DIF cannot be detected accurately enough and thus the item scores are invalidated as a surrogate for the abilities. The joint analysis of the re-weighted sample leads in the first step to a maximal value of the statistic of 0.234 attained at the intent of persons who answered the question F26: "Which internet company took over the media group Time Warner? - AOL." correctly and had a score value between 17 and 37. The minimum of the test statistic in the first step was -0.333 attained at an intent containing the 5 questions

F12: "Which form of government is associated with the French King Louis XIV? - Absolutism."

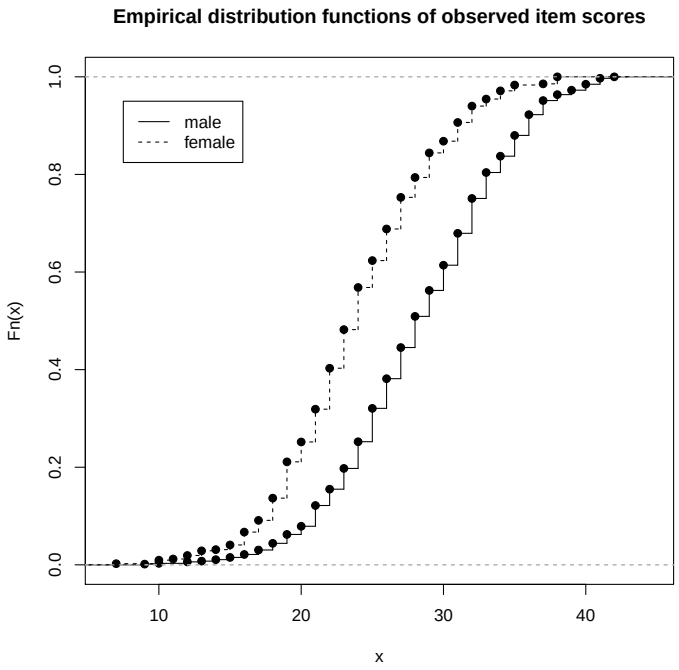

Figure 7: Empirical distribution of the item scores for the female and the male group in the subsample of the general knowledge quiz "Studentenpisa" ([SPIEGEL Online, 2009]).

F33: "What is the name of the bestselling novel by Daniel Kehlmann? - Die Vermessung der Welt (Measuring The World)."

F35: "In which city is this building located? - Paris."

F40: "What is also termed Trisomy 21? - Down syndrome."

F43: "Which kind of bird is this? - Blackbird."

and score values between 16 and 35. After excluding questions F26, F12, F33 F35, F40 and F43 for matching, in a second step, additionally the two questions

F34: "Which city is the setting for the novel 'Buddenbrooks'? - Lübeck." and

F36: "Which one of the following operas is not by Mozart? - Aida."

were excluded for matching. In the third step, the procedure stopped with a maximal final statistic D^+ of 0.241 attained for the intent containing $F26$ and (modified) score values between 16 and 29. The minimal value D^- was -0.290 attained for the intent containing questions $F12$, $F33$, $F35$, $F40$ and $F43$ and (modified) score values between 12 and 30. Thus, altogether, questions F12, F26 , F33, F34, F35, F36, F40 and F43 showed DIF. (The result was statistically significant in the sense that a bootstrapp sample of 10

sample yielded a distribution of the absolute value of the test statistic with mean 0.21 and standard deviation 0.01.)

6.3 Guided taming of concept extents in cognitive diagnosis models

In this section, we would like to illustrate a little bit, how one can tame a formal context in a more guided way in the context of cognitive diagnosis models. For illustration, we use a subsample of the *Trends in International Mathematics and Science Study* (TIMSS) of the year 2007. This study is an international assessment of the mathematics and science knowledge of students, that was firstly conducted in 1995 and has been administered every four years thereafter by the International Association for the Evaluation of Educational Achievement (IEA). It analyses math- and science knowledge of 4th and 8th grade students. We use here a subsample provided in the R-package `CDM`[43], consisting of 698 Austrian students (4th grade) answering a set of 25 math questions (dataset `data.timss07.G4.lee`). Since not all students answered all 25 questions, we restrict here the analysis to that 344 students that answered all questions. The 25 questions were the same as that used in Lee et al. [2011]. The package also provides the Q-matrix and the description of the skills used in Lee et al. [2011]. We will use this small subsample to illustrate the guided taming procedure by comparing it to the non-guided taming procedure described in section 5.3.2. The formal context $\mathbb{K}_0 = (\{g_1, \ldots, g_{344}\}, \{m_1, \ldots, m_{25}\}, I)$ has a Vapnik-Chervonenkis dimension of 14 and consists of 255712 formal concepts. Because of the small cardinality of the concept lattice, we can explicitly compute the closure system $\mathfrak{B}_1(\mathbb{K}_0)$ of all extents and thus we will analyze the taming process not w.r.t. the V.C.-dimension, but w.r.t. the cardinalities of the tamed closure systems $\mathfrak{B}_1(\tilde{\mathbb{K}})$. The data set contains also information about gender, so we will analyze differences w.r.t. gender. Figure 8 shows the value of the test statistic for the actually observed data in dependence on the cardinality of the tamed closure system for both the guided taming and the non-guided taming. One can see that, as expected, the statistic increases with increasing cardinality of the closure system. The general pictures for the non-guided and the guided taming are very similar. For the guided taming, the smallest closure system that is obtained by enforcing all valid implications of the idealized response pattern space, has a size of 127, which is much higher than the smallest possible closure system of size 2, obtainable by the strongest possible non-guided taming. Figure 9 shows the p-value one would obtain if one would do a statistical test. (Here, we did resampling with 1000 resamples to compute the p-values.) One can see, that for comparable sizes of the closure system, the guided taming procedure generally has lower p-values. One could speculate here, that the guided taming tends to exclude mainly sets that are statistically not so important in the sense that they play no crucial role w.r.t. differences between male and female participants. If one assumes that in the actually observed data set there are clear differences between male and female participants,

[43] See Robitzsch et al. [2016] for an introduction to the package `CDM`.

then it seemingly appears here, that the guided taming leads to a smart reduction of the size of the closure system that actually reduces the variability of the statistic under H_0 without reducing the test statistic under H_1, too much.

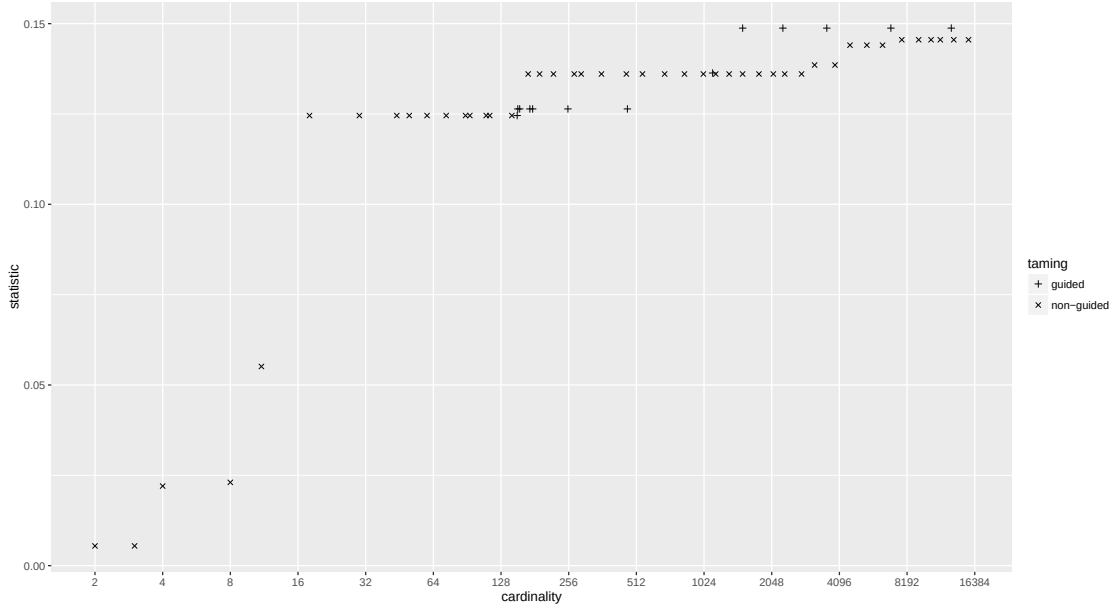

Figure 8: Value of the tamed test statistic for different cardinalities of the tamed closure system, both for the non-guided and the guided taming.

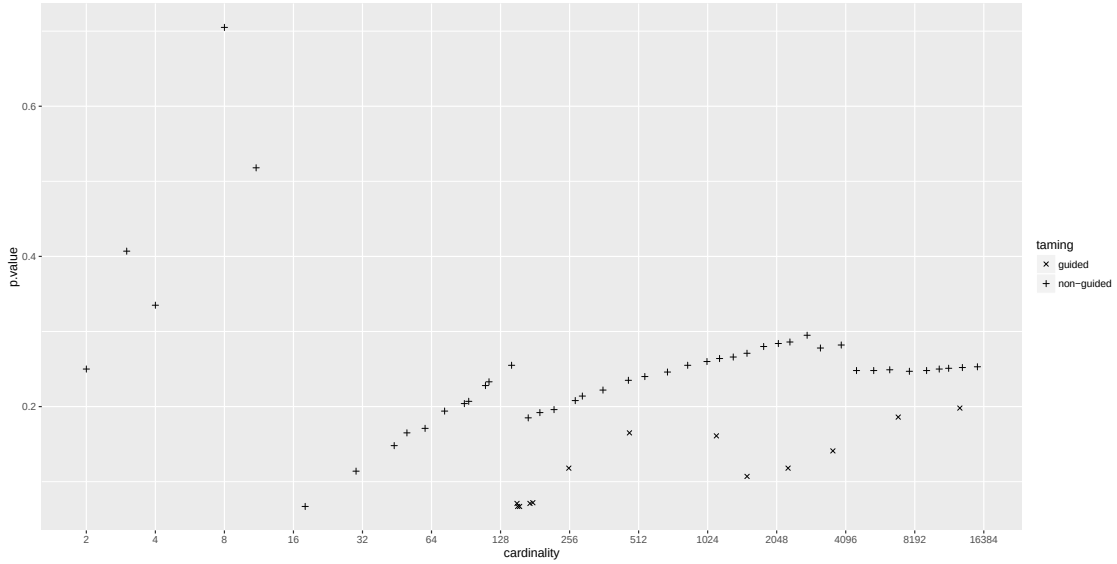

Figure 9: Obtained p-values of the test statistic for the actually observed data for different sizes of the tamed closure system, both for the non-guided and the guided procedure.

6.4 Convex sets: A geometrical generalization of the Kolmogorov-Smirnov test

Finally, we want to illustrate that for small data sizes the generalization of the Kolmogorov-Smirnov test for analyzing spatial differences between subpopulations in spatial statistics as indicated in Section 4 is also practically applicable. We use here the data set `quercusvm` which is a subsample of a larger data set analyzed in Laskurain [2008] and available in the R Package *ecespa* ([de la Cruz Rot, 2008]). This data set consists of 100 data points representing the locations of alive and dead oak trees (Quercus robur) in a secondary wood in Urkiola Natural Park (Basque country, north of Spain). The data are depicted in Figure 10. We can now compute that convex sets where the maximal and the minimal

Figure 10: Locations of altogether 100 alive and dead oak tress (Quercus robur) in a secondary wood in Urkiola Natural Park (Basque country, north of Spain).

differences in proportion of alive and dead oak trees is attained. Figure 11 shows the

results. The blue convex set is the set, where the difference is maximal (37%): In the

Figure 11: Difference in proportions of alive and dead oak trees. Blue: maximal difference of 39% (more alive than dead trees). Red: minimal difference of −37% (more dead than alive trees).

blue convex area we have a proportion of 61% alive, but only a proportion of 24% dead trees. The red convex set is the set, where the difference is minimal (−39%): In the red convex area there are 15% alive and 54% dead trees. Based on 1000 resamples, one gets an approximate p-value of 0.83, so the differences or not statistically significant. Note that also the Cramér von Mises type test proposed by Syrjala [1996] and also a classical generalization of the Kolmogorov-Smirnov test, where one only looks at rectangular areas are both non-significant. (The p-values of the tests, computed with the function `syrjala` from the R Package `ecespa` are approximately 0.84 and 0.67, respectively.) Compared to Syrjalas test, the test based on convex sets has the advantage that it is somehow better

interpretable because one can actually see, in which areas the differences in proportion are maximal or minimal.

A further modification of the test is also possible: Since the convex sets are described by formal implications and one explicitly models these implications by imposing corresponding inequality constraints in the binary program, one has the flexibility to impose not all, but only some implications. One natural way to select implications to include would be to include only implications where the data points of the premise and the conclusion are not too far away from each other. This would lead to some kind of a localization method and the associated closure system would get larger, which means more flexibility in detecting non-convex distributional features but a generally higher V.C.-dimension. We shortly illustrate this modification by imposing only implications where the distances between the points of the premises and the conclusions is not greater than $40m$. Figure 12 shows the non-convex sets where the maximal (blue) and the minimal (red) differences in the proportions of alive and dead oak trees is attained.

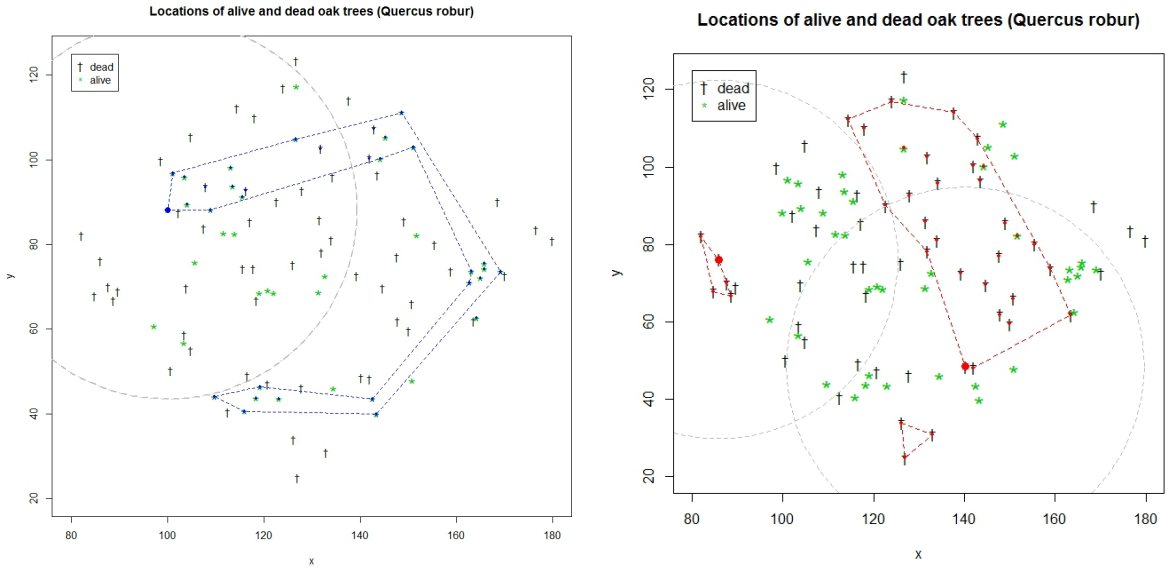

Figure 12: Non-convex sets where the maximal (blue) and the minimal (red) difference in proportions between alive and dead oak trees is attained for the modified method where only implications, where the distance between the points of the premises and the conclusions in not greater than $40m$, are used. The dashed circles around the highlighted blue and red points have a radius of $40m$ to get an impression about which implications were actually included.

For the modified version we got a maximal value of 58% and a minimal value of 50% for the difference of the proportions with an approximate p-value of 0.17.

While the computations for this data set could be done quickly enough, for larger data sets, the method quickly becomes intractable. The data set analyzed in Syrjala [1996] containing 327 spatial measuring points was already very hard to analyze, to compute the test statistic, the mixed integer solver Gurobi (Gu, Rothberg, and Bixby [2012]) took a few days to solve the binary program. Similarly to the analysis in Syrjala [1996], the result w.r.t. differences between male and female cods was not statistically significant. To assess the statistical significance of our test statistic, we did not need to do resampling, which would actually be very time demanding. Instead we could rely on the fact that for a value of 0.06 that we observed for our test statistic, still the more classical Kolmogorov-Smirnov type test statistic that only looks at rectangular areas would not be statistically significant.

However, for dealing with the computational issue, one can use the technique of attribute exploration for formal contexts: One can firstly look at the formal context $\mathbb{K} := (G, M, I)$ where G is the set of all rectangular areas, M is the set of all spatial measuring points and gIm iff measuring point m lies in the rectangular area g. The resulting closure system of all concept intents is then the set of all sets of measuring points lying in some rectangular area, which is a smaller closure system than the system of all convex sets of measuring points and in which thus more formal implications are valid. Note that despite this, a base of all implications of this smaller closure system can be given as

$$\{\{p, q\} \mapsto [p, q] \mid p, q \in M, [p, q] \supsetneq \{p, q\}\},$$

where $[p, q] := \{r \in M \mid r_1 \in [\min\{p_1, q_1\}, \max\{p_1, q_1\}] \,\&\, r_2 \in [\min\{p_2, q_2\}, \max\{p_2, q_2\}]\}$. Compared to the base for general convex sets, this base has only $\mathcal{O}(n^2)$ implications and is thus far more easy to handle.

Now, during the computation of all valid implications of $\mathbb{K}$, in the spirit of attribute exploration, one can check for every currently generated implication, if it is also approximately true with some confidence c in the context $\tilde{\mathbb{K}} = (\tilde{G}, M, \tilde{I})$, where $\tilde{G}$ is the set of all half-spaces generated by two points of M and $g\tilde{I}m$ means that measuring point m lies in the half-space g. The intents of this context are exactly all convex areas of measuring points. If the currently generated implication is also true in the larger closure system of all convex areas, then one would treat it as valid, otherwise one would provide a convex half-space $g \in \tilde{G}$ as a counterexample.

With this procedure one would generate a closure system that is larger than the system of all rectangular areas and smaller than the closure system of all convex areas and the confidence level c regulates the size of the resulting closure system and the size of the implication base.

Thus, with this modification, we have some "scalable" method for spatial statistics. (Of course, with the drawback that now the result of the method is dependent on the choice of the coordinate system.)

7 Conclusion

In this paper we analyzed the problem of detecting stochastic dominance as a prototypical example of optimizing a linear function on a closure system. Compared to the general case, for stochastic dominance, the integrality constraints of the underlying binary program could be dropped which helped in making the problem more tractable. For general closure systems the binary programs are more difficult to solve, but we managed to solve them in our concrete cases of application. Note that we did not explicitly incorporate knowledge about the underlying closure system into the mixed integer solver we used. It seems that one can make the computations far more efficient by using for example knowledge about valid formal implications of the underlying closure system. If one knows that the certain formal implications are valid, then one can possibly use this knowledge to explicitly prune the search space in the branch and cut algorithm of the mixed integer solver.

The solved binary programs and the associated test statistics treated in this paper could be understood as some Kolmogorov-Smirnov type generalizations. This motivates the question if also other generalizations like weighted Kolmogorov-Smirnov type or Anderson-Darling type tests are computational tractable. Actually, it seems to be not too difficult to compute such variants of a test statistic: Firstly, one can impose one additional constraint into the underlying program that demands that the sets one is optimizing over contain at least (or at most, or exactly) an amount c of overall probability mass. Secondly, one can do the constrained optimization for every possible amount c and can then aggregate the optimal values for different c for example to

$$\sup_c \sup_{\substack{m: \\ \langle w^x + w^y, m\rangle \geq c}} \langle w^x - w^y, m\rangle \cdot \psi(c),$$

where ψ is some appropriately chosen weighting function.

All in all, it seems that the optimization of linear functions on closure systems has a broad range of possible applications and thus deserves further research.

References

R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216. ACM, 1993.

A. Albano. The implication logic of (n,k)-extremal lattices. In K. Bertet, D. Borchmann, P. Cellier, and S. Ferré, editors, *Formal Concept Analysis: 14th International Conference, ICFCA 2017, Rennes, France, June 13-16, 2017, Proceedings*, pages 39–55. Springer, 2017a.

A. Albano. *Polynomial growth of concept lattices, canonical bases and generators: extremal set theory in formal concept analysis*. PhD thesis, Technische Universität Dresden,

2017b. URL http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-226980. accessed 19.08.2017.

A. Albano and B. Chornomaz. Why concept lattices are large: extremal theory for generators, concepts, and VC-dimension. *International Journal of General Systems*, pages 1–18, 2017.

S. Alkire, J. Foster, S. Seth, J. Roche, and M. Santos. *Multidimensional Poverty Measurement and Analysis*. Oxford University Press, 2015.

M. Anthony, N. Biggs, and J. Shawe-Taylor. The learnability of formal concepts. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 246–257. Morgan Kaufmann Publishers Inc., 1990a.

M. Anthony, N. Biggs, and J. Shawe-Taylor. *Learnability and Formal Concept Analysis*. University of London. Royal Holloway and Bedford New College. Department of Computer Science, 1990b.

C. Arndt, R. Distante, M. A. Hussain, L. P. Østerdal, P. L. Huong, and M. Ibraimo. Ordinal welfare comparisons with multiple discrete indicators: A first order dominance approach and application to child poverty. *World Development*, 40(11):2290–2301, 2012.

C. Arndt, M. A. Hussain, V. Salvucci, F. Tarp, and L. P. Østerdal. Advancing small area estimation. Technical report, WIDER Working Paper, 2013. URL http://hdl.handle.net/10419/80908. accessed 28.08.2017.

C. Arndt, N. Siersbæk, and L. P. Østerdal. Multidimensional first-order dominance comparisons of population wellbeing. Technical report, WIDER Working Paper, 2015. URL http://hdl.handle.net/10419/129471.

M. M. Babbar. Distributions of solutions of a set of linear equations (with an application to linear programming). *Journal of the American Statistical Association*, 50(271):854–869, 1955.

G. F. Barrett and S. G. Donald. Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104, 2003.

Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In J. Lloyd, V. Dahl, U. Fuhrbach, M. Kerber, K.-K. Lau, C. Palamidessi, and L. M. Pereira, editors, *Computational Logic-CL 2000*, pages 972–986. Springer, 2000.

G. Birkhoff. Rings of sets. *Duke Mathematical Journal*, 3(3):443–454, 1937. doi: 10.1215/S0012-7094-37-00334-X.

L. Bottou. In hindsight: Doklady akademii nauk SSSR, 181(4), 1968. In B. Schölkopf, Z. Luo, and V. Vovk, editors, *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 3–5. Springer, 2013.

B. Chornomaz. Counting extremal lattices. working paper or preprint, 2015. URL `https://hal.archives-ouvertes.fr/hal-01175633`.

B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2002.

O. Davidov and S. Peddada. Testing for the multivariate stochastic order among ordered experimental groups with application to dose–response studies. *Biometrics*, 69(4):982–990, 2013.

M. de la Cruz Rot. Metodos para analizar datos puntuales. In F. T. Maestre, A. Escudero, and A. Bonet, editors, *Introduccion al Analisis Espacial de Datos en Ecologia y Ciencias Ambientales: Metodos y Aplicaciones.*, chapter 3, pages 76–127. Asociacion Espanola de Ecologia Terrestre, Universidad Rey Juan Carlos and Caja de Ahorros del Mediterraneo, 2008. URL `https://cran.r-project.org/web/packages/ecespa/index.html`.

J.-P. Doignon and J.-C. Falmagne. *Knowledge Spaces*. Springer, 2012.

B. Dushnik and E. W. Miller. Partially ordered sets. *American Journal of Mathematics*, 63(3):600–610, 1941.

B. Ganter and R. Wille. *Formal concept analysis: mathematical foundations*. Springer, 2012.

GESIS - Leibniz - Institut fur Sozialwissenschaften. Allbus compact (2015): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, 2015. URL `https://www.gesis.org/allbus/inhalte-suche/studienprofile-1980-bis-2016/2014/`. GESIS Datenarchiv, Köln. ZA5241 Datenfile Version 1.1.0.

Z. Gu, E. Rothberg, and R. Bixby. Gurobi optimizer reference manual. 2012. URL `http://www.gurobi.com`. accessed 19.08.2017.

E. H. Haertel. Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4):301–321, 1989.

G. Hansel and J. Troallic. Mesures marginales et théorème de Ford-Fulkerson. *Probability Theory and Related Fields*, 43(3):245–251, 1978.

J. Heller, L. Stefanutti, P. Anselmi, and E. Robusto. On the link between cognitive diagnostic models and knowledge space theory. *Psychometrika*, 80(4):995–1019, Dec 2015.

P. W. Holland, D. T. Thayer, H. Wainer, and H. Braun. Differential item performance and the Mantel-Haenszel procedure. *Test validity*, pages 129–145, 1988.

J. E. Hopcroft and R. M. Karp. A n5/2 algorithm for maximum matchings in bipartite. In *Switching and Automata Theory, 1971., 12th Annual Symposium on*, pages 122–125. IEEE, 1971.

B. W. Junker and K. Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272, 2001.

T. Kamae, U. Krengel, and G. L. O'Brien. Stochastic inequalities on partially ordered spaces. *The Annals of Probability*, 5(6):899–912, 1977.

T. Kuosmanen. Efficient diversification according to stochastic dominance criteria. *Management Science*, 50(10):1390–1406, 2004.

L. Lakhal and G. Stumme. Efficient mining of association rules based on formal concept analysis. *Formal concept analysis*, 3626:180–195, 2005.

N. Laskurain. *Dinámica espacio-temporal de un bosque secundario en el Parque Natural de Urkiola (Bizkaia)*. PhD thesis, Universidad del País Vasco/Euskal Herriko Unibertsitatea, 2008.

Y.-S. Lee, Y. S. Park, and D. Taylan. A cognitive diagnostic modeling of attribute mastery in massachusetts, minnesota, and the us national sample using the timss 2007. *International Journal of Testing*, 11(2):144–177, 2011.

E. Lehmann. Ordered families of distributions. *Ann Math Stat*, 26:399–419, 1955.

M. Leshno and H. Levy. Stochastic dominance and medical decision making. *Health Care Management Science*, 7(3):207–215, 2004.

D. Levhari, J. Paroush, and B. Peleg. Efficiency analysis for multivariate distributions. *The Review of Economic Studies*, 42(1):87–91, 1975.

H. Levy. *Stochastic dominance: Investment decision making under uncertainty*. Springer, 2015.

H. Levy and M. Levy. Experimental test of the prospect theory value function: A stochastic dominance approach. *Organizational Behavior and Human Decision Processes*, 89(2): 1058 – 1081, 2002.

T. Makhalova and S. O. Kuznetsov. On overfitting of classifiers making a lattice. In K. Bertet, D. Borchmann, P. Cellier, and S. Ferré, editors, *Formal Concept Analysis: 14th International Conference, ICFCA 2017*, pages 184–197. Springer, 2017.

K. Mosler and M. Scarsini. Some Theory of Stochastic Dominance. In K. Mosler and M. Scarsini, editors, *Stochastic orders and decision under risk*, pages 203–212. Institute of Mathematical Statistics, 1991.

A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks*. Wiley, 2002.

S. J. Osterlind and H. T. Everson. *Differential Item Functioning*. Sage Publications, 2009.

G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–248, 1991.

J. W. Pratt and J. D. Gibbons. *Concepts of Nonparametric Theory*. Springer, 2012.

A. Prêkopa. On the probability distribution of the optimum of a random linear program. *SIAM Journal on Control*, 4(1):211–222, 1966.

C. Preston. A generalization of the FKG inequalities. *Communications in Mathematical Physics*, 36(3):233–241, 1974.

T. M. Range and L. P. Østerdal. Checking bivariate first order dominance. Technical report, Discussion Papers on Business and Economics, 2013.

A. Robitzsch, T. Kiefer, A. C. George, and A. Uenlue. *CDM: Cognitive Diagnosis Modeling*, 2016. URL `http://CRAN.R-project.org/package=CDM`. R package version 4.8-0.

A. Rusch and R. Wille. Knowledge spaces and formal concept analysis. In H.-H. Bock and W. Polasek, editors, *Data Analysis and Information Systems: Statistical and Conceptual Approaches Proceedings of the 19th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Basel*, pages 427–436. Springer, 1996.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

H. Scheiblechner. A unified nonparametric IRT model for d-dimensional psychological test data (d-ISOP). *Psychometrika*, 72(1):43, 2007.

J. K. Sengupta, G. Tintner, and B. Morrison. Stochastic linear programming with applications to economic models. *Economica*, 30(119):262–276, 1963.

S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

SPIEGEL Online. Studentenpisa - Alle fragen, alle Antworten, 2009. URL `http://www.spiegel.de/unispiegel/studium/0,1518,620101,00.html`. In German. accessed 18.08.2017.

V. Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

C. Strobl, J. Kopf, and A. Zeileis. Rasch trees: A new method for detecting differential item functioning in the rasch model. *Psychometrika*, 80(2):289–316, 2015.

S. E. Syrjala. A statistical test for a difference between the spatial distributions of two populations. *Ecology*, 77(1):75–80, 1996.

F. Tarp and L. P. Østerdal. Multivariate discrete first order stochastic dominance. Discussion Papers 07-23, University of Copenhagen. Department of Economics, 2007. URL `http://EconPapers.repec.org/RePEc:kud:kuiedp:0723`.

K. K. Tatsuoka. *Analysis of errors in fraction addition and subtraction problems*. National Institute of Education, Washington, D.C., 1984.

S. Trepte and M. Verbeet. *Allgemeinbildung in Deutschland: Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test*. VS Verlag für Sozialwissenschaften, 2010.

W. T. Trotter. *Combinatorics and Partially Ordered Sets: Dimension Theory*, volume 6. JHU Press, 2001.

G. Tutz and M. Berger. Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, 81(3):727–750, 2016.

G. Tutz and G. Schauberger. A penalty approach to differential item functioning in Rasch models. *Psychometrika*, 80(1):21–43, 2015.

UNESCO Institute for Statistics (UIS). *International Standard Classification of Education: ISCED 2011*. UIS, Montreal, Quebec, 2012.

V. N. Vapnik and S. Kotz. *Estimation of Dependences Based on Empirical Data*. Springer, 1982.

N. Vayatis and R. Azencott. Distribution-dependent vapnik-chervonenkis bounds. In P. Fischer and H. U. Simon, editors, *European Conference on Computational Learning Theory*, pages 230–240. Springer, 1999.

R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In *Ordered Sets*, pages 445–470. Springer, 1982.

R. Zwick. When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, 15(3):185–197, 1990.