Mathematische Strukturen* für Studierende der Statistik

Georg Schollmeyer

Wintersemester 2019/20

„Die Ordnungstheorie ist, wie viele Teilgebiete der Mathematik, einfach und anspruchsvoll zugleich, abstrakt und angewandt, anschaulich und unvorstellbar.“ [Bernhard Ganter]

*Der Titel dieser Veranstaltung mag ein wenig irreführend sein, es geht hier beispielsweise nicht um die Theorie algebraischer Strukturen, sondern eher um die Anwendung ordnungs- und verbandstheoretischer Überlegungen in der Sprache der formalen Begriffsanalyse zur Behandlung von Fragestellungen im Kontext (deskriptiver wie schließender) statistischer Datenanalyse.

Inhaltsverzeichnis

1 Inzidenzstrukturen

Definition 1.1 (Inzidenzstruktur, duale Inzidenzstruktur)
Eine **Inzidenzstruktur** ist ein Tripel (G, M, I), wobei G und M beliebige Mengen sind[1] und $I \subseteq G \times M$ eine Teilmenge des kartesischen Produktes von G und M ist. Wenn G und M aus dem Kontext heraus klar sind, dann wird die Menge I auch als (binäre) **Relation** (zischen G und M) bezeichnet. Für ein $g \in G$ und ein $m \in M$ schreiben wir gIm falls $(g, m) \in I$ und sagen, dass g mit m inzidiert. Wenn $G \cap M = \emptyset$ und somit keine Verwechslungsgefahr besteht, dann sagt man im Fall gIm auch, dass m mit g inzidiert. Die zu (G, M, I) **duale Inzidenzstruktur** ist gegeben durch (M, G, I^∂) mit $I^\partial := \{(m, g) \in M \times G \mid gIm\}$.

Beispiele:

a) **(synthetische) Geometrie**:

G	...	Menge von Punkten
M	...	Menge von Geraden
gIm	...	wird interpretiert als: Punkt g liegt auf Gerade m.

b) **Rasch-Modell**:

G	...	Menge von Personen
M	...	Menge von Aufgaben
gIm	...	wird interpretiert als: Person g hat Aufgabe m erfolgreich gelöst.

c) **BTL-Modell (Bradley-Terry-Luce-Modell)**:

G	...	Menge von Items
$M = G$	...	Menge von Items
gIm	...	wird interpretiert als: Item g wird gegenüber Item m bevorzugt.

d) **Formale Begriffsanalyse**:

G	...	Menge von **G**egenständen
M	...	Menge von **M**erkmalen
gIm	...	wird interpretiert als: Gegenstand g besitzt Merkmal m.

Bemerkung 1.1. Oft erfüllt die Relation I bestimmte Struktureigenschaften: Beispielsweise existiert in Beispiel a) für je zwei verschiedene Punkte g_1 und g_2 genau eine Gerade m, die mit beiden Punkten inzidiert. Das Konzept der Inzidenzstruktur ist jedoch für beliebige Relationen sinnvoll. Beispielsweise in Anwendungen der formalen Begriffsanalyse innerhalb einer relationalen Datenanalyse ist die Inzidenz I im Zweifelsfalle völlig kontingent.

Bemerkung 1.2. Manchmal ist es hilfreich, die **duale Inzidenzstruktur** (M, G, I^∂) mit $I^\partial := \{(m, g) \mid m \in M, g \in G : gIm\}$ zu betrachten: In Beispiel a) würde $mI^\partial g$ bedeuten, dass die Gerade m den Punkt g enthält. Für zwei Geraden m_1 und m_2 existiert dann, sofern sie nicht

[1] Einige Autoren wie beispielsweise Beutelspacher [1982] fordern zusätzlich noch $G \cap M = \emptyset$, was wir hier aber nicht tun.

parallel sind, genau ein Punkt, der mit beiden Geraden inzidiert, d.h., die duale Inzidenz hat sehr ähnliche Struktureigenschaften wie die ursprüngliche Inzidenz. Man kann nun für parallele Geraden künstliche (virtuelle) Fernpunkte[2] einführen, in denen sich diese Geraden schneiden. Führt man zusätzlich noch eine Ferngerade ein, die mit allen Fernpunkten (und nur mit diesen) inzidiert, dann ist die duale Inzidenz noch strukturähnlicher (isomorph?) zur usrprünglichen Inzidenz. (Verbunden mit einem zusätzlichen Reichhaltigkeitsaxiom führt dies auf den Begriff der projektiven Ebene, siehe `http://en.wikipedia.org/wiki/Projective_plane`.)

Definition 1.2 (Homogene Inzidenzstruktur, Eigenschaften homogener Inzidenzstrukturen)
Eine Inzidenzstruktur (G, M, I) mit $G = M =: X$ heißt homogen und wird mit (X, I) abgekürzt. Eine **homogene** Inzidenzstruktur heißt

i) **reflexiv**, falls $\forall x \in X : xIx$;

ii) **transitiv**, falls $\forall x, y, z \in X : xIy \ \& \ yIz \Longrightarrow xIz$;

iii) **symmetrisch**, falls $\forall x, y \in X : xIy \Longrightarrow yIx$;

iv) **antisymmetrisch**, falls $\forall x, y \in X : xIy \ \& \ yIx \Longrightarrow x = y$;

v) **total**, **linear** oder **konnex**, falls $\forall x, y \in X : xIy$ oder yIx.

Eine homogene Inzidenzstrukur (X, I) heißt

a) **Äquivalenzrelation**, falls (X, I) reflexiv, transitiv und symmetrisch ist;

b) **Ordnungsrelation** oder **geordnete Menge**, falls (X, I) reflexiv, transitiv und antisymmetrisch ist;

c) **Quasiordnung**, **Präordnung**, oder auch **quasigeordnete Menge** bzw. **prägeordnete Menge**, falls (X, I) reflexiv und transitiv ist.

Eine Teilmenge $M \subseteq X$ einer geordneten Menge $(X, \leq)$ heißt

d) **Kette**, falls $(M, \leq \cap M \times M)$ eine total geordnete Menge ist.

e) **Antikette**, falls beliebige verschiedene Elemente $x, y \in M$ unvergleichbar sind, d.h., dass weder $x \leq y$, noch $y \leq x$ gilt.

Bemerkung 1.3. Da für reflexive Inzidenzstrukturen (X, I) die Menge X aus der Relation I via $X = \{x \mid (x, x) \in I\}$ ableitbar ist, werden in diesem Fall die entsprechenden Begriffe jeweils sowohl für das Paar (X, I), als auch für die eigentliche Inzidenz I verwendet.

Bemerkung 1.4. Für geordnete Mengen $(X, \leq)$ wird die duale Relation zu $\leq$ mit $\geq$ bezeichnet.

[2] Vergleiche auch Hilberts „Methode der idealen Elemente“, diskutiert z.B. in Tapp, Christian: An den Grenzen des Endlichen: Das Hilbertprogramm im Kontext von Formalismus und Finitismus, siehe `https://edoc.ub.uni-muenchen.de/6523/1/Tapp_Christian.pdf` oder `http://www.springer.com/de/book/9783642296536`

1.1 Äquivalenzrelationen

Beispiel 1. $(M, =)$ mit M einer beliebigen Menge und der üblichen Gleichheitsrelation $=$.

Beispiel 2. Sei M eine Menge von Zufallsvariablen mit dem gleichen Bildmessraum $(\Omega', \mathcal{A}')$. Definiere für zwei Zufallsvariablen $X : (\Omega, \mathcal{A}, P) \longrightarrow (\Omega', \mathcal{A}')$ und $Y : (\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P}) \longrightarrow (\Omega', \mathcal{A}')$ die Relation

$$X \stackrel{d}{=} Y :\iff \forall A \in \mathcal{A}' : P(X \in A) = \tilde{P}(Y \in A).$$

Dann ist $\stackrel{d}{=}$ eine Äquivalenzrelation auf M.

Beispiel 3. Sei M eine Menge von Zufallsvariablen mit gleichem Bildmessraum und mit gleichem Urbildmaßraum $(\Omega, \mathcal{A}, P)$. Dann ist $\stackrel{P-f.s.}{=} := \{(X, Y) \in M \times M \mid P(X = Y) = 1\}$ eine Äquivalenzrelation auf M.

Beispiel 4. Sei Ω eine Menge und und $F \subseteq \mathbb{R}^\Omega$ eine Menge von Funktionen von Ω nach $\mathbb{R}$. Dann ist die **Ununterscheidbarkeitsrelation** $\sim_F := \{(\omega_1, \omega_2) \mid \omega_1, \omega_2 \in \Omega, \forall f \in F : f(\omega_1) = f(\omega_2)\}$ eine Äquivalenzrelation.

Bemerkung 1.5. Zum Paar (Ω, F) ist auf natürliche Weise das **duale Paar** (F, Ω') mit $\Omega' := \{\omega' \mid \omega \in \Omega\}$ einer Menge von **Auswertungsfunktionalen** von F nach $\mathbb{R}$ definiert, wobei ein Auswertungsfunktional ω' hier in natürlicher Weise durch $\omega' : F \longrightarrow \mathbb{R} : f \mapsto f(\omega)$ gegeben ist. Durch das duale Paar (F, Ω') wird dann die duale Ununterscheidbarkeitsrelation $\sim_{\Omega'} := \{(f, g) \mid f, g \in F, \forall \omega \in \Omega : \underbrace{\omega(f)}_{=f(\omega)} = \underbrace{\omega(g)}_{=g(\omega)}\}$, die nichts anderes, als die übliche Gleichheitsrelation für Funktionen ist, induziert. Betrachtet man allgemeiner für $A \subseteq \Omega$ das Paar (F, A') mit $A' := \{\omega' \mid \omega \in A\}$, so erhält man als Äquivalenzrelation die Gleichheit von Funktionen auf der Einschränkung A.

1.2 Ordnungsrelationen

Beispiel 5. Die Reellen Zahlen, ausgestattet mit der üblichen Ordnung, bilden eine (total) geordnete Menge.

Beispiel 6. (Paradebeispiel)
$(2^M, \subseteq)$ ... Die Menge aller Teilmengen einer Menge M mit der Teilmengenrelation bildet eine geordnete Menge. Ebenso bildet jedes Mengensystem $\mathcal{S} \subseteq 2^M$ zusammen mit der Teilmengenrelation eine geordnete Menge. **Fakt:** Man kann sich jede geordnete Menge als ein solches $\mathcal{S}$ mit geeignet gewähltem Ω vorstellen, siehe später.

Beispiel 7. Die natürlichen Zahlen $\mathbb{N}$ bilden zusammen mit der Teilerrelation $m|n :\iff \exists c \in \mathbb{N} : m \cdot c = n$ eine geordnete Menge.

Beispiel 8. Die **natürliche Relation** der natürlichen Zahlen:

$$m \leq n :\iff \exists c \in \mathbb{N} : m + c = n$$

ist eine (lineare) Ordnungsrelation auf $\mathbb{N}$.

Beispiel 9. Sei A eine Menge und $F \subseteq B^A$ eine Menge von Abbildungen von A nach B, wobei B mit der Ordnung $\leq$ ausgestattet sei.
Dann ist

$$\leq^A := \{(f,g) \in F \times F \mid \forall x \in A : f(x) \leq g(x)\}$$

eine Ordnungsrelation auf F. Für $C \subsetneq A$ ist

$$\leq^C := \{(f,g) \in F \times F \mid \forall x \in C : f(x) \leq g(x)\}$$

im Allgemeinen lediglich eine Quasiordnung.

Beispiel 10. Sei M eine Menge von Zufallsvariablen mit gleichem Urbildmaßraum und mit gleichem Bildmessraum, welcher zusätzlich mit der Ordnungsrelation $\leq$ ausgestattet sei. Dann ist $\leq_{P-f.s.} := \{(X,Y) \in M \times M \mid P(X \leq Y) = 1\}$ eine Äquivalenzrelation auf M.

Beispiel 11. Sei M eine Menge und K eine Menge von (homogenen) Relationen auf M. Dann ist durch $(K, \subseteq)$ eine geordnete Menge gegeben.

1.3 Darstellung geordneter Mengen

Definition 1.3 (Nachbarschaftsrelation, Diagonale, Hassegraph)
Die einer Ordnung $(X, \leq)$ zugeordnete **Nachbarschaftsrelation** $\lessdot$ ist gegeben durch

$$\lessdot := \{(x,y) \mid x,y \in X, x \neq y, x \leq y, \forall z \in X : x \leq z \leq y \Longrightarrow z = x \text{ oder } z = y\}.$$

Die Nachbarschaftsrelation $\lessdot$ charakterisiert eine endliche Ordnungsrelation eindeutig via

$$\begin{aligned} \leq &= H(\lessdot) := \bigcap \{ \sqsubseteq \mid \sqsubseteq \text{ Ordnungsrelation auf } X \text{ mit } \sqsubseteq \supseteq \lessdot \} \\ &= \{(x,y) \mid x,y \in X, \exists z_1 \lessdot z_2 \ldots \lessdot z_k : x = z_1, y = z_k\} \cup \Delta_X, \end{aligned}$$

wobei hier $\Delta_X := \{(x,x) \mid x \in X\}$ die sogenannte **Diagonale** auf X ist. Der **Hassegraph** einer endlichen Ordnungsrelation besteht aus Punkten, die die Elemente der Menge X darstellen, und aus Kanten zwischen diesen Punkten. Dabei liegt ein Punkt x immer oberhalb eines Punktes y falls $y \leq x$. Gilt zusätzlich $y \lessdot x$, so sind x und y zusätzlich durch eine Kante verbunden, d.h., es werden gewissermaßen nur die für die Darstellung wichtigen Kanten, die sich nicht automatisch durch Reflexivität und Transitivität ergegeben, dargestellt.

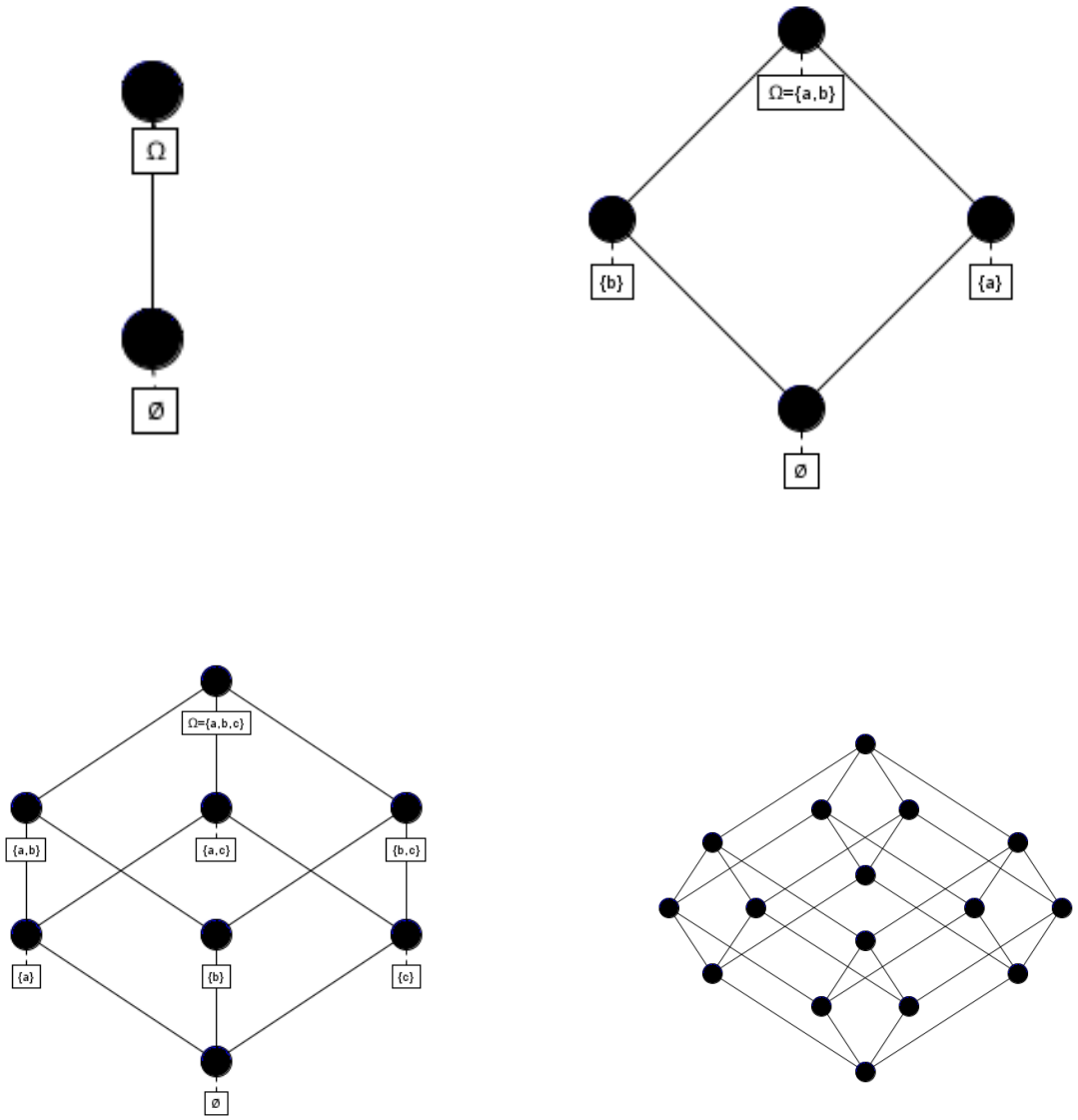


Abbildung 1: Hasse-Diagramme der Potenzmengenverbände einer 1-, 2-, 3- bzw. 4-elementigen Grundmenge Ω.

Satz 1.4 (Satz von Dushnik und Miller)
Jede Ordnungsrelation $\leq$ auf einer Menge X ist darstellbar als der Schnitt aller ihrer totalen Erweiterungen:

$$\leq \quad = \quad \bigcap \{R \quad | \quad R \text{ lineare Ordnung auf } X \quad \& \quad R \supseteq \leq\}. \tag{1}$$

Obiger Satz bedeutet, dass man sich jede geordnete Menge $(X, \leq)$ als Schnitt einer Menge M von totalen Ordnungen vorstellen kann, $x \leq y$ kann man sich also auch als „x ist in jeder Hinsicht (im Sinne von „bezüglich jeder totalen Ordnung $R \in M$“) kleinergleich y“ vorstellen. Dies ist natürlich zunächst nur in einem strukurellen Sinne, nicht in einem inhaltlichen Sinne zu verstehen, die totalen Relationen aus M müssen nicht notwendigerweise eine natürliche Interpretation besitzen.

Bemerkung 1.6. Es ist nicht selbstverständlich, dass der Schnitt in (1) nichtleer ist. Für endliche geordnete Mengen kann man aus $\leq$ Schritt für Schritt eine totale lineare Erweiterung von $\leq$ erhalten, indem man folgendermaßen vorgeht: Für $i = 1, 2, \ldots$:

i) Setze $\sqsubseteq_1 := \leq$.

ii) Falls $\sqsubseteq_i$ noch zwei unvergleichbare Elemente b, c enthält, dann setze

$$\sqsubseteq_{i+1} := \sqsubseteq_i \cup \{(x, y) \mid x \sqsubseteq_i b \;\&\; c \sqsubseteq_i y\}.$$

Dies ist in der Tat eine Ordnungsrelation auf X, die die Relation $\leq$ erweitert und insbesondere das vorher unvergleichbare Paar (b, c) enthält.

iii) Falls $\sqsubseteq_{i+1}$ noch weitere unvergleichbare Elemente b', c' enthält, fahre analog zu Schritt (ii) fort usw., bis letztendlich die Relation $\sqsubseteq_n$ keine unvergleichbaren Paare mehr enthält, also eine lineare Ordnung ist und nach Konstruktion auch die ursprüngliche Relation $\leq$ erweitert.

Für eine endliche geordnete Menge kommt man damit in der Tat nach endlich vielen Schritten zum Ziel, denn es kann ja nur maximal endlich viele unvergleichbare Elemente geben. Für unendliche geordnete Mengen müsste man zusätzlich noch das Zornsche Lemma bemühen.

Bemerkung 1.7. Man hätte in (1) auch lineare Präordnung anstelle von lineare Ordnung schreiben können, denn jede lineare Ordnung ist eine lineare Präordnung und zu jeder linearen Präordnung $\lesssim$, die $\leq$ erweitert kann man mit einer beliebigen linearen Ordnung L, die $\leq$ erweitert, via

$$\sqsubseteq = \{(x, y) \mid x \lesssim y \text{ oder } (x \sim y \;\&\; xLy)\}$$

eine lineare Ordnung konstruieren, die dasselbe wie $\lesssim$ leistet.

Die Darstellung geordneter Mengen als Schnitte von totalen Ordnungen ermöglicht es, einen Begriff der „Komplexität“ einer geordneten Menge zu entwickeln: Je mehr totale Ordnungen nötig sind, um eine Ordnung darzustellen, desto komplexer ist die Ordnung. Da die Darstellung einer geordneten Menge als Schnitt von totalen Ordnungen nicht eindeutig ist, bietet sich folgende Definition der **Ordnungdimension** an:

Definition 1.5
Die **Ordnungdimension** $\operatorname{odim}((X, \leq))$ einer geordneten Menge $(X, \leq)$ ist die minimale Anzahl von totalen Ordnungen, die nötig sind, um die Ordnung $(X, \leq)$ als Schnitt dieser Ordnungen darzustellen.

Bemerkung 1.8. Nach obigen Bemerkungen hätte man in der Definition der Ordnungsdimension auch mit linearen Präordnungen anstelle von linearen ordnungen arbeiten können und würde zur gleichen Begriffsbildung gelangen.

Beispiel 12. Für Ω endlich gilt:

$$\operatorname{odim}((2^\Omega, \subseteq)) = \operatorname{odim}((\mathbb{R}^\Omega, \leq^\Omega)) = |\Omega|.$$

(Hier ist $f \leq^\Omega g :\iff \forall \omega \in \Omega : f(\omega) \leq g(\omega)$ für $f, g \in \mathbb{R}^\Omega$.)

Beweisidee (für $(2^\Omega, \subseteq)$):

Betrachte die Projektionspräordnungen $\pi_\omega := \{(A, B) \mid \mathbb{1}_A(\omega) \leq \mathbb{1}_B(\omega)\}$. Dann gilt offensichtlich

$$\subseteq = \bigcap_{\omega \in \Omega} \pi_\omega,$$

was bedeutet, dass die Ordnungdimension höchstens $|\Omega|$ sein kann. Um zu sehen, dass die Ordnungsdimension nicht kleiner als $|\Omega|$ sein kann, betrachte alle einelementigen Mengen und deren Komplemente: Jede beliebige einelementige Menge A is unvergleichbar mit ihrem Komplement A^c. Deshalb muss es in der Darstellung von $\subseteq$ als Schnitt von linearen Ordnungen für jedes einelementige A eine totale lineare Erweiterung L_A geben mit $(A, A^c) \notin L_A$. Das sind insgesamt also $|\Omega|$ lineare Erweiterungen von denen wir nur noch zeigen müssen, dass sie alle paarweise verschieden sind. Seien dazu verschiedene einelementige Mengen A und B mit $L_A = L_B =: L$ gegeben. Dann würde folgen

$$(A^c, A) \in L \;\&\; (B^c, B) \in L.$$

Dies kann aber nicht sein, denn da L die Relation $\subseteq$ erweitert, folgt wegen $A \subseteq B^c$ mit der Transitivität von L die Relation $(A^c, B) \in L$, was im Widerspruch zu $B \subseteq A^c$ steht, also müssen L_A und L_B für verschiedene einelementige Mengen A und B in der Tat verschieden sein.

1.4 Verbände

Definition 1.6 (obere, untere Schranke, maximales, minimales Element, größtes, kleinstes Element)
Ein Element $s \in X$ einer geordneten Menge $(X, \leq)$ heißt **obere (untere) Schranke** einer Menge $S \subseteq X$, falls für alle Elemente $x \in S$ die Relation $s \geq x$ (bzw. $s \leq x$) erfüllt ist. Eine Menge $S \subseteq X$ besitzt ein **größtes (kleinstes) Element** s, falls $s \in S$ und $\forall x \in S : s \geq x$ (bzw. $\forall x \in S : s \leq x$)
gilt. Weiterhin heißt ein Element $s \in S$ **maximales (minimales) Element** von S„ falls es kein weiteres Element $y \in S$ gibt mit $y > s$ (bzw. $y < s$). Das größte (kleinste) Element einer geordneten Menge ist immer eindeutig bestimmt.

Definition 1.7 (Verband, vollständiger Verband)
Eine geordnete Menge $(X, \leq)$ heißt **Verband**, falls für zwei beliebige Elemente $x, y \in X$ die Menge der oberen Schranken von $\{x, y\}$ immer ein kleinstes Element (genannt kleinste obere Schranke oder **Supremum** von $\{x, y\}$) besitzt und die Menge der unteren Schranken von $\{x, y\}$ ein größtes Element (genannt größte untere Schranke oder **Infimum** von $\{x, y\}$) besitzt. Falls beliebige Teilmengen $S \subseteq X$ (einschließlich der leeren Menge) immer eine kleinste obere Schranke und eine größte untere Schranke besitzen, so wird $(X, \leq)$ als **vollständiger Verband** bezeichnet. Die kleinste obere Schranke einer Menge S wird mit $\bigvee S$ und die größte untere Schranke wird mit $\bigwedge S$ bezeichnet.

Bemerkung 1.9. Das Infimum und das Supremum sind immer eindeutig bestimmt. Das Infimum der leeren Menge ist die größte untere Schranke von nichts, also das größte Element schlechtin, denn alle Elemente sind untere Schranken der leeren Menge. Das Supremum der leeren Menge ist die kleinste obere Schranke von nichts, also das kleinste Element schlechthin. Also besitzt jeder vollständige Verband immer ein grösstes und ein kleinstes Element. Weiterhin lässt sich in einem vollständigen Verband die Eigenschaft, obere/untere Schranke zu sein, charakterisieren durch

$$x \text{ ist obere Schranke von } S \iff x \geq \bigvee S \text{ bzw.} \tag{2}$$

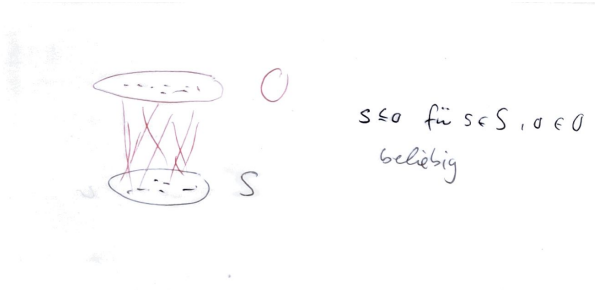
$$x \text{ ist untere Schranke von } S \iff x \leq \bigwedge S. \tag{3}$$

Satz 1.8 (Infimum induziert Supremum)
In einer geordneten Menge $(X, \leq)$, in der beliebige Infima existieren, besitzt jede nach oben beschränkte Menge $S \subseteq X$ auch ein Supremum, nämlich

$$\bigvee S = \bigwedge \{x \mid x \text{ obere Schranke von } S\}.$$

Dabei wird S als nach oben beschränkt bezeichnet, falls S wenigstens eine obere Schranke besitzt.

Beweis.
Setze $O := \{x \mid x \text{ ist oberee Schranke von } S\}$ und $z := \bigwedge O$.



Dann ist z eine obere Schranke von S, denn für beliebiges $s \in S$ gilt:

$$\begin{aligned} z \geq s &\iff \bigwedge O \geq s \\ &\iff s \text{ ist untere Schranke von } O \\ &\iff \forall o \in O : s \leq o \end{aligned}$$

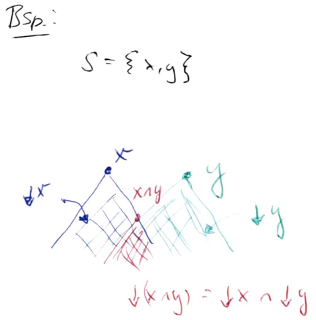
und jedes $o \in O$ war eine obere Schranke von S. Außerdem ist z die kleinste obere Schranke von S, denn für jede weitere obere Schranke $\tilde{z}$ von S gilt $\tilde{z} \in O$ und damit $z \leq \tilde{z}$, denn z war die größte untere Schranke von O und damit insbesondere eine untere Schranke von O. □

Satz 1.9 (Freie Darstellung des Infimums)
Für eine geordnete Menge $(X, \leq)$ definiere zunächst für $x \in X$:

$$\begin{aligned} \downarrow x &:= \{y \in X \mid y \leq x\} \text{ bzw.} \\ \uparrow x &:= \{y \in X \mid y \geq x\}. \end{aligned}$$

In einem vollständigen Verband $(X, \leq)$ gilt für beliebige Teilmengen $S \subseteq X$ immer

$$\bigcap_{s \in S} \downarrow s = \downarrow \bigwedge S.$$

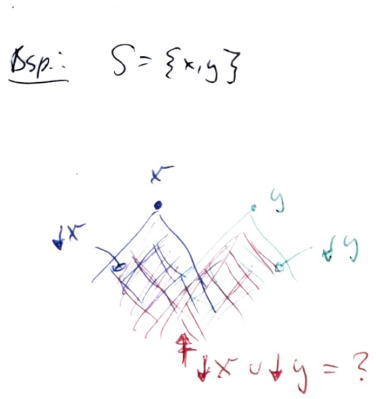


Beweis.
Links wie rechts steht nichts anderes als die Menge aller unteren Schranken von S. □

Bemerkung 1.10. Die linke Seite obiger Gleichung ist immer auch für geordnete Mengen, die keinen vollständigen Verband bilden, definiert, was später noch relevant sein wird.

Frage: Was ist mit

$$A := \bigcup_{s \in S} \downarrow s?$$



A ist im Allgemeinen nicht darstellbar als $A = \downarrow x$ für ein $x \in X$. Allerdings ist A immer noch eine sogenannte Unterhalbmenge, d.h., es gilt die Implikation

$$a \in A, y \leq a \Longrightarrow y \in A,$$

vergleiche dazu auch den späteren Abschnitt zur stochastischen Dominanz.

1.5 Abbildungen zwischen geordneten Mengen

Definition 1.10 ((starker) Homomorphismus)
Seien $(X, \leq_X)$ und $(Y, \leq_Y)$ geordnete Mengen. Eine Abbildung $f : (X, \leq_X) \longrightarrow (Y, \leq_Y)$ heißt

- (ordnungstheoretischer) **Homomorphismus** bzw. auch **isoton**, falls

$$\forall x, x' \in X : x \leq_X x' \Longrightarrow f(x) \leq_Y f(x').$$

- **starker** (ordnungstheoretischer) **Homomorphismus**, falls

$$\forall x, x' \in X : x \leq_X x' \iff f(x) \leq_Y f(x').$$

Bemerkung 1.11. Ein starker Homomorphismus ist bereits injektiv, denn es gilt $f(x) = f(x') \Longrightarrow f(x) \leq_Y f(x')$ & $f(x') \leq_Y f(x) \Longrightarrow x \leq_X x'$ & $x' \leq_X x' \Longrightarrow x = x'$.

Beispiel 13.

$$f : (X, \leq) \longrightarrow (2^X, \subseteq) : x \mapsto \downarrow x := \{y \in X \mid y \leq x\}$$

ist ein starker (ordnungstheoretischer) Homomorphismus, denn es gilt

$$x \leq x' \iff \downarrow x \subseteq \downarrow x'.$$

Beweis.

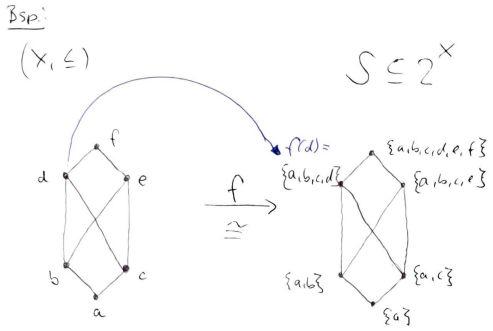
"$\Longrightarrow$": Sei $x \leq x'$. Dann ist $\downarrow x \subseteq \downarrow x'$, denn für jedes $y \in \downarrow x$ gilt $y \leq x$ und wegen $x \leq x'$ auch $y \leq x'$ (denn $\leq$ ist transitiv), woraus unmittelbar $y \in \downarrow x'$ folgt.

"$\Longleftarrow$": Sei $\downarrow x \subseteq \downarrow x'$. Aus $x \in \downarrow x$ folgt $x \in \downarrow x'$ und somit $x \leq x'$.

□

Satz 1.11 (Mengentheoretischer Darstellungssatz)
Jede geordnete Menge $(X, \leq)$ ist darstellbar als Unterstruktur eines Potenzmengenverbandes, d.h., es gibt einen bijektiven starken Homomorphismus zwischen $(X, \leq)$ und $(\mathcal{S}, \subseteq)$ für geeignet gewähltes $\mathcal{S} \subseteq 2^\Omega$ und geeignet gewähltes Ω.

Beweis.
Der Satz folgt unmittelbar aus der Betrachtung der Abbildung f aus Beispiel 13 und der Setzung $\Omega := X$ sowie $\mathcal{S} := \mathrm{im}(f) := \{f(x) \mid x \in X\} = \{\downarrow x \mid x \in X\}$. □



Definition 1.12 (Adjunktion)
Es seien $(X, \leq_X)$ und $(Y, \leq_Y)$ zwei geordnete Mengen sowie $f : (X, \leq_X) \longrightarrow (Y, \leq_Y)$ und $g : (Y, \leq_Y) \longrightarrow (X, \leq_X)$ Abbildungen. Dann bilden f und g eine **Adjunktion** zwischen $(X, \leq_X)$ und $(Y, \leq_Y)$, falls für alle $x \in X$ und alle $y \in$Y die Äquivalenz

$$f(x) \leq_Y y \iff x \leq_X g(y)$$

gilt.

Beispiel 14. Für eine geordnete Menge $(X, \leq)$ definiere

$$\Phi : (2^X, \subseteq) \longrightarrow (2^X, \supseteq) : A \mapsto \text{ Menge aller unteren Schranken von } A \text{ und}$$
$$\Psi : (2^X, \supseteq) \longrightarrow (2^X, \subseteq) : B \mapsto \text{ Menge aller oberen Schranken von } B.$$

Dann bilden Φ und Ψ eine Adjunktion zwischen $(2^X, \subseteq)$ und $(2^X, \supseteq)$, denn es gilt:

$$\underbrace{\Phi(A) \supseteq B}_{b \leq a \text{ für alle } a \in A \text{ und alle } b \in B.} \iff \underbrace{A \subseteq \Psi(B)}_{a \geq b \text{ für alle } b \in B \text{ und alle } a \in A.} .$$

Bemerkung 1.12. Bei obiger Konstruktion war es überhaupt nicht von Belang, dass $(X, \leq)$ eine homogene Ordnungsrelation ist. Deshalb noch das folgende weitere Beispiel einer beliebigen, im Allgemeinen nicht-homogenen Inzidenzstruktur, die keinerlei besondere Struktureigenschaften zu tragen braucht:

Beispiel 15 (Formale Begriffsanalyse). Gegeben der formale Kontext $\mathbb{K} := (G, M, I)$ mit G einer Menge von Gegenständen, M einer Menge von Merkmalen und $I \subseteq G \times M$ einer binären Relation zwischen G und M mit der Interpretaton $gIm \iff$ Gegenstand g besitzt Merkmal m. Betrachte dann die Abildungen

$$\Phi : (2^M, \subseteq) \longrightarrow (2^G, \supseteq) : \quad B \mapsto B' := \underbrace{\{g \in G \mid \forall m \in B : gIm\}}_{\text{Menge aller Gegenstände, die alle Merkmale aus } B \text{ besitzen.}} \text{ und}$$

$$\Psi : (2^G, \supseteq) \longrightarrow (2^M, \subseteq) : \quad A \mapsto A' := \underbrace{\{m \in M \mid \forall g \in A : gIm\}}_{\text{Menge aller den Gegenständen aus } A \text{ gemeinsamer Merkmale.}}.$$

Beobachtung: Für $A \subseteq G$ und $B \subseteq M$ gilt:

$$\underbrace{\underbrace{\Phi(B)}_{\substack{\text{Menga aller Gegenstände,}\\ \text{die alle Merkmale aus } B\\ \text{besitzen.}}} \supseteq A}_{\substack{\text{Alle Gegenstände aus } A \text{ haben}\\ \text{alle Merkmale aus } B.}} \iff \underbrace{B \subseteq \underbrace{\Psi(A)}_{\substack{\text{Menge aller den Gegen-}\\ \text{ständen aus } A \text{ gemeinsa-}\\ \text{mer Merkmale.}}}}_{\substack{\text{Alle Merkmale aus } B \text{ werden}\\ \text{von allen Gegenständen aus } A\\ \text{getragen. Anders ausgedrückt:}\\ \text{Alle Gegenstände aus } A \text{ haben}\\ \text{alle Merkmale aus } B.}} .$$

Also: Das Paar (Φ, Ψ) bildet eine Adjunktion zwischen den geordneten Mengen $(2^M, \subseteq)$ und $(2^G, \supseteq)$.

Satz 1.13 (Adjunktion induziert Hüllenoperator)
Für eine Adjunktion (f, g) zwischen $(X, \leq_X)$ und $(Y, \leq_Y)$ gilt:

i) f und g und damit auch $g \circ f$ und $f \circ g$ sind isoton.

ii) $g \circ f$ ist ein Hüllenoperator (siehe gleich);

iii) $f \circ g$ ist ein Kernoperator (siehe ebenfalls gleich);

iv) $f \circ g \circ f = f$;

v) $g \circ f \circ g = g$;

vi) f erhält existierende Suprema, d.h., für Mengen $S \subseteq X$ mit (in $(X, \leq_X)$) existierendem Supremum $\bigvee S$ besitzt die Menge $f(S) := \{f(s) \mid s \in S\}$ (in $(Y, \leq_Y)$) ein Supremum und es gilt $f(\bigvee S) = \bigvee f(S)$;

vii) g erhält existierende Infima.

Beweis.
Vergleiche Übung 1. □

Definition 1.14 (Hüllenoperator, Kernoperator)
Sei $(X, \leq)$ eine geordnete Menge. Eine Abbildung $h : (X, \leq) \longrightarrow (X, \leq)$ heißt (ordnungstheoretischer) **Hüllenoperator**, falls sie folgende Bedingungen erfüllt:

(Extensivität): $\forall x \in X : x \leq h(x)$

(Isotonie): $\forall x, y \in X : x \leq y \Longrightarrow h(x) \leq h(y)$

(Idempotenz): $\forall x \in X : h(h(x)) = h(x)$.

Eine isotone, idempotente Abbildung $k : (X, \leq) \longrightarrow (X, \leq)$, die intensiv ist (intensiv heißt $\forall x \in X : k(x) \leq x$), wird als Kernoperator bezeichnet.

Bemerkung 1.13. Ein Kernoperator auf $(X, \leq)$ ist also nichts anderes als ein Hüllenoperator auf der dualen Ordnungsrelation $(X, \geq)$.

Bemerkung 1.14. Ist $X = 2^\Omega$ für eine Menge Ω und $\leq = \subseteq$ die gewöhnliche Mengeninklusion, so spricht man auch von einem **mengentheoretischen Hüllenoperator**.

Definition 1.15 (Hüllensystem)
Sei $(X, \leq)$ ein vollständiger Verband. Eine Menge $\mathcal{H} \subseteq X$, die das größte Element $\top$ von $(X, \leq)$ enthält und die abgeschlossen unter beliebiger Infmumbildung ist (, d.h., für jede Teilmenge $Y \subseteq \mathcal{H}$ ist das Infimum $\bigwedge Y$ ebenfalls wieder aus $\mathcal{H}$,) heißt (ordnungstheoretisches) **Hüllensystem**.

Satz 1.16 (Zusammenhang Hüllenoperator Hüllensystem)
Sei $(X, \leq)$ ein vollständiger Verband. Dann gilt:

i) Jeder Hüllenoperator $h : (X, \leq) \longrightarrow (X, \leq)$ induziert das Hüllensystem
$$\mathcal{H}_h := \{h(x) \mid x \in X\}.$$

ii) Jedes Hüllensystem $\mathcal{H} \subseteq X$ induziert den Hüllenoperator
$$h_\mathcal{H} : (X, \leq) \longrightarrow (X, \leq) : x \mapsto \bigwedge \{y \mid y \in \mathcal{H}, y \geq x\}.$$

iii) Darüber hinaus gelten die Identitäten $h_{\mathcal{H}_h} = h$ und $\mathcal{H}_{h_\mathcal{H}} = \mathcal{H}$.

Beispiel 16 (Hüllenoperatoren und Hüllensysteme).

1. Sei $\Omega = \mathbb{R}^2$ und
$$h : 2^\Omega \longrightarrow 2^\Omega : A \mapsto \text{co}\,(A)$$
der Operator, der jeder Menge ihre konvexe Hülle zuordnet. Dieser Operator ist ein Hüllenoperator. Weiterhin ist das System $\mathcal{H}$ aller konvexen Teilmengen von $\mathbb{R}^2$ ein Hüllensystem und es gilt:
$$h(A) = \bigcap \{B \mid B \in \mathcal{H} \;\&\; B \supseteq A\}.$$

2. Maßtheorie: Sei Ω Grundraum und
$$\sigma : 2^{2^\Omega} \longrightarrow 2^{2^\Omega} : \mathcal{E} \mapsto \sigma(\mathcal{E}) := \bigcap \{\mathcal{A} \mid \mathcal{A} \supseteq \mathcal{E}, \mathcal{A} \text{ ist } \sigma\text{-Algebra über } \Omega\}$$
der Operator, der jedem Mengensystem $\mathcal{E}$ die von diesem Mengensystem erzeugte σ-Algebra zuordnet. Dieser Operator ist ein Hüllenoperator, denn das Mengensystem $\mathcal{H} = \{\mathcal{A} \mid \mathcal{A} \text{ ist } \sigma - \text{Algebra über } \Omega\}$ ist vollständig schnittstabil und enthält das größte Element 2^Ω.

3. Transitive Hülle: Sei X eine Menge und $\mathcal{H} = \{R \mid R \text{ ist transitive Relation auf } X\}$. Betrachte die geordnete Menge $(2^{X \times X}, \subseteq)$. Dann ist $(2^{X \times X}, \subseteq)$ ein vollstaändiger Verband in dem das Infimum durch den Mengenschnitt und das Supremum durch die Mengenvereinigung gegeben ist. Beobachtung: Das größte Element $R = X \times X$ ist transitiv und beliebige Schnitte von transitiven Relationen sind wieder transitiv. Also bildet $\mathcal{H}$ ein Hüllensystem und damit ist die Abbildung
$$h : (2^{X \times X}, \subseteq) \longrightarrow (2^{X \times X}, \subseteq) : R \mapsto \bigcap \{S \mid S \in \mathcal{H} \;\&\; S \supseteq R\}$$
ein Hüllenoperator.

4. Sei (X, d) ein vollständiger metrischer Raum. Dann ist die Abbildung

$$H : 2^X \longrightarrow 2^X : A \mapsto \underbrace{\bar{A}}_{\text{Abschluss der Menge } A = \text{ kleinste abgeschlossene Menge, die } A \text{ enthält.}}$$

ein Hüllenoperator bzw. die Menge aller abgeschlossenen Mengen eines vollständigen metrischen Raumes bildet ein Hülllensystem.

5. Sei $(X, \leq)$ eine geordnete Menge. Betrachte die Menge $\mathcal{U}((X, \leq))$ aller Unterhalbmengen[3] von $(X, \leq)$. Dann ist $\mathcal{U}((X, \leq))$ ein Hüllensystem. Beweis:

   a) Das größte Element X ist eine Unterhalbmenge, denn für jedes $y \in X$ gilt schon $y \in X$.

   b) $\mathcal{U}((X, \leq))$ ist abgeschlossen unter beliebigen Schnitten: Sei dazu $(Y_i)_{i \in I}$ eine Klasse von Unterhalbmengen. Betrachte $Z := \bigcap\{Y_i \mid i \in I\}$. Dann ist Z wieder eine Unterhalbmenge: Sei dazu $x \in Z$ & $y \leq x$. Dann gilt für alle $i \in I$, dass $x \in Y_i$ und somit $y \in Y_i$, denn alle Y_i waren Unterhalbmengen und $y \leq x$. Somit ist y auch wieder aus Z.

Bemerkung 1.15. Das System aller Unterhalbmengen einer geordneten Menge ist sogar ein vollständiger Mengenring. (Ein vollständiger Mengenring ist eine Familie von Mengen, die unter beliebigen Vereinigungen und Schnitten abgeschlossen ist. (Vergleiche auch den späteren Abschnitt zur stochastischen Dominanz.) Der zugehörige Hüllenoperator kann konkret angegeben werden als:

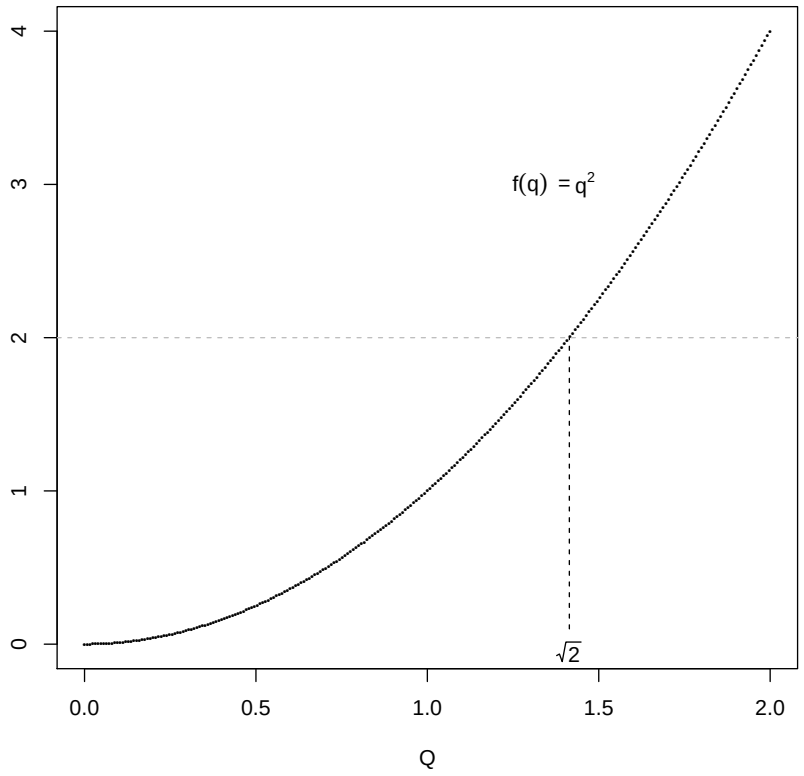
$$h : 2^X \longrightarrow 2^X : A \mapsto \downarrow A := \bigcup_{a \in A} \downarrow a.$$

[3]Eine Menge $A \subseteq X$ heißt Unterhalbmenge, falls $\forall x, y \in X : x \in A \;\&\; y \leq x \Longrightarrow y \in A$ gilt.

1.6 Vorbereitungen zur Dedekind-MacNeille-Vervollständigung und zur Formalen Begriffsanalyse

Frage: Was, wenn $(X, \leq)$ kein vollständiger Verband ist?

Antwort: Hilberts Methode der idealen Elemente, konkret im Stile der Dedekindschen Schnitte: Man erinnere sich an die in der ersten Vorlesung angesprochene Konstruktion der reellen Zahlen als Vervollständigung der rationalen Zahlen:



In $\mathbb{Q}$ gibt es bekanntlich keine Zahl, deren Quadrat gleich 2 ist, aber: Man kann sich eine rationale Zahl q aus ordnungstheoretischer Sicht auch so vorstellen:
Jedes „reale“ Element $q \in \mathbb{Q}$ „zerlegt“ $\mathbb{Q}$ in $\downarrow q$ und $\uparrow q$. (Beachte: $\downarrow q \cap \uparrow q = \{q\}$ und $\downarrow q \cup \uparrow q = \mathbb{Q}$.) Von der „Zahl“ $\sqrt{2}$ „wissen“ die rationalen Zahlen zwar nichts, aber die „Dedekindschen Schnitte“

$$U := \{q \in \mathbb{Q} \mid q^2 \leq 2\} \text{ und}$$
$$O := \{q \in \mathbb{Q} \mid q^2 \geq 2\}$$

„charakterisieren“ in gewissem Sinne (sozusagen als „ideale Elemente“, die man einfach hinzufügen könnte, sofern sie nicht schon existierten,) die irrationale „Lücke“ $\sqrt{2}$.

Frage: Funktioniert diese Idee auch für eine beliebige geordnete Menge?
Antwort: Ja.

Sei also $(X, \leq)$ eine beliebige geordnete Menge. Dann „zerlegt“ jedes „reale“ Element $x \in X$ die Menge X in 3 Teile:

1. alles unterhab von x: $U_x :=\downarrow x$,
2. alles oberhalb von x: $O_x :=\uparrow x$
3. und den Rest $X\backslash(U_x \cup O_x)$.

Das charakteristische an dieser „Zerlegung“ ist, dass

a) $u \leq o$ für beliebige $u \in U_x$ und $o \in O_x$ gilt, und dass

b) die Mengen U_x und O_x maximal bezüglich dieser Eigenschaft sind.

Man könnte jetzt „ideale“ Elemente analog definieren: Ein „ideales“ Element ist ein Paar (U, O) von Teilmengen von X mit $u \leq o$ für belieibige $u \in U$ und $o \in O$ und die Mengen U und O sind maximal bezüglich dieser Eigesnchaft, d.h., es gibt kein $u \in X\backslash U$ mit $u \leq O$ für alle $o \in O$ und es gibt kein $o \in X\backslash O$ mit $u \leq o$ für alle $u \in U$. Das ist im Prinzip alles. Die Menge aller idealen Elemente werden wir zukünftig mit $\mathfrak{B}((X, X, \leq))$ bezeichnen.

Frage: Wie bekommt man alle idealen Elemente?

Ein (natürlich recht ineffizientes) Vorgehen:

i) Starte mit einem beliebigen Kandidaten $\tilde{O} \subseteq X$ für „das Oberhalb“.

ii) Definiere U als

$$\begin{aligned} U :&= \{x \in X \mid x \leq o \text{ für alle } o \in \tilde{O}\} = \Phi(\tilde{O}) \\ &= \bigcap_{o \in \tilde{O}} \downarrow o \text{ (Vgl. auch „freie Darstellung“ des Infimums).} \end{aligned}$$

U ist dann wirklich schon ein zulässiges „Unterhalb“.

iii) Und dann? Definiere

$$\begin{aligned} O :&= \{x \in X \mid x \geq u \text{ für alle } u \in U\} = \Psi(U) = (\Psi \circ \Phi)(\tilde{O}) \\ &= \bigcap_{u \in U} \uparrow u \text{ (ist i.A. } \supsetneq \bigcup_{o \in \tilde{O}} \uparrow o). \end{aligned}$$

Dann ist jedes so konstruierte Paar

$$\Big(\Phi(\tilde{O}), (\Psi \circ \Phi)(\tilde{O})\Big)$$

in der Tat ein ideales Element und jedes ideale Element entsteht so. Alternativ kann mann auch mit einem beliebigen Kandidaten $\tilde{U}$ für das „Unterhalb“ starten und dann

$$\Big((\Phi \circ \Psi)(\tilde{U}), \Psi(\tilde{U})\Big)$$

betrachten.

Bemerkungen: Ein Paar (U, O) ist nach Konstruktion ein „ideales“ Element genau dann, wenn

$$\Phi(O) = U \text{ und}$$
$$\Psi(U) = O$$

gilt. Dies macht klar, warum man sich jedes „ideale“ Element (U, O) als nach obiger Konstruktion entstanden vorstellen kann, man beginne in der Konstruktion einfach schon mit dem fertigen „Oberhalb“ O, also man setze $\tilde{O} := O$. Dann ist

$$\Big(\Phi(\tilde{O}), (\Psi \circ \Phi)(\tilde{O})\Big) = (\Phi(O), (\Psi \circ \Phi)(O)) = (U, \Psi(\Phi(O))) = (U, \Psi(U)) = (U, O).$$

Beachte auch, dass Φ und Ψ eine Adjunktion bilden und dass deshalb

$$\Phi \circ \Psi \circ \Phi = \Phi \text{ und}$$
$$\Psi \circ \Phi \circ \Psi = \Psi$$

gilt, d.h., weitere Anwendungen von Φ bzw. Ψ in obiger Konstruktion von U und O würden dem „Oberhalb“ und dem „Unterhalb“ nichts weiter hinzufügen.

Beispiele:
a) Wenn $(X, \leq)$ bereits ein vollständiger Verband ist:
Dann gilt für $\tilde{O} \subseteq X$ beliebig:

$$\begin{aligned} U = \Phi(\tilde{O}) &= \bigcap_{o \in \tilde{O}} \downarrow o = \downarrow \bigwedge \tilde{O} \text{ und} \\ O = \Psi(U) &= \Psi\left(\downarrow \bigwedge \tilde{O}\right) \\ &= \text{ Menge der oberen Schranken von } \downarrow \bigwedge \tilde{O} \\ &= \uparrow \bigwedge \tilde{O}, \end{aligned}$$

denn die oberen Schranken von $\downarrow \bigwedge \tilde{O}$ sind genau die Elemente, die größergleich $\bigwedge \tilde{O}$ sind.
Also:

$$U = \downarrow x$$
$$O = \uparrow x$$

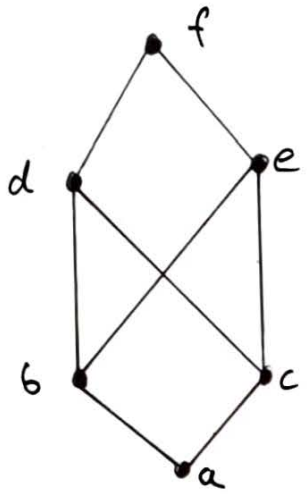
für $x := \bigwedge \tilde{O}$ $(= \bigwedge O = \bigvee U)$.

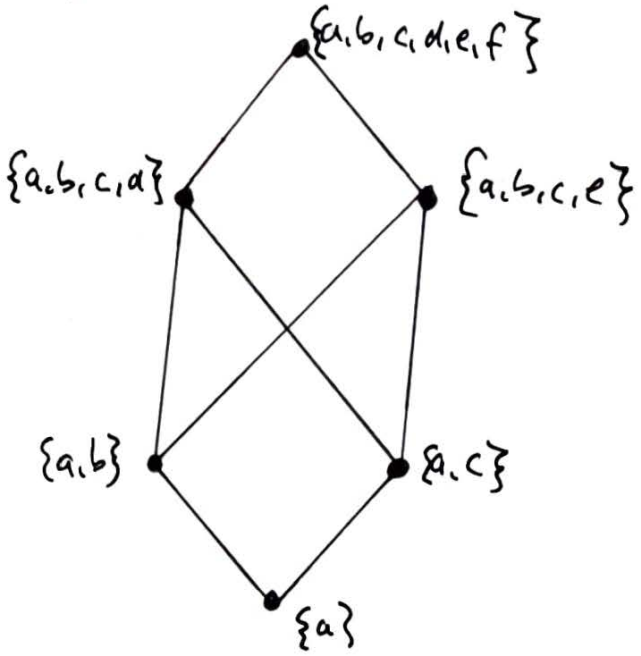
Also: Alle „idealen“ Elemente (U, O) haben eine „Entsprechung“ mit „realen“ Elementen x via $(\downarrow x, \uparrow x)$ mit $x = \bigwedge O = \bigvee U$.

b) Wenn $(X, \leq)$ kein vollständiger Verband ist:

$(X, \leq)$ $\xrightarrow{f:}$ $(\mathrm{im}(f), \subseteq)$

$f(x) = \downarrow x$





bzw.:

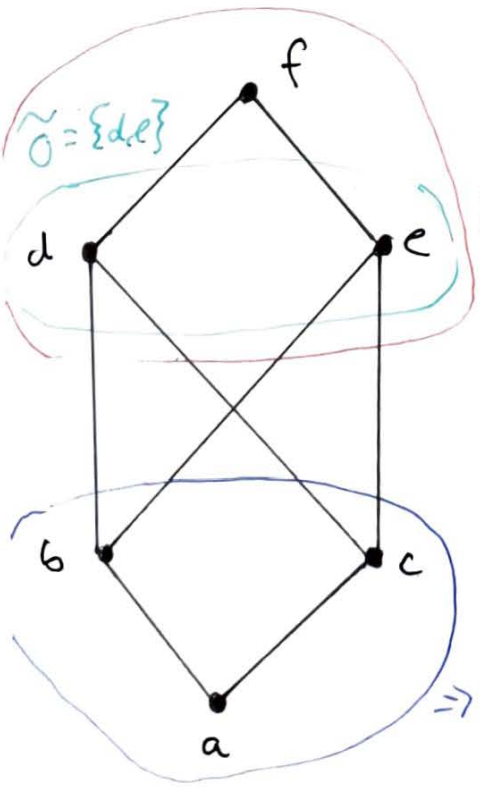
$(X, \leq)$ $\xrightarrow{g:}$ $(\mathrm{im}(g), \sqsubseteq)$

$g(x) = (\downarrow x, \uparrow x)$

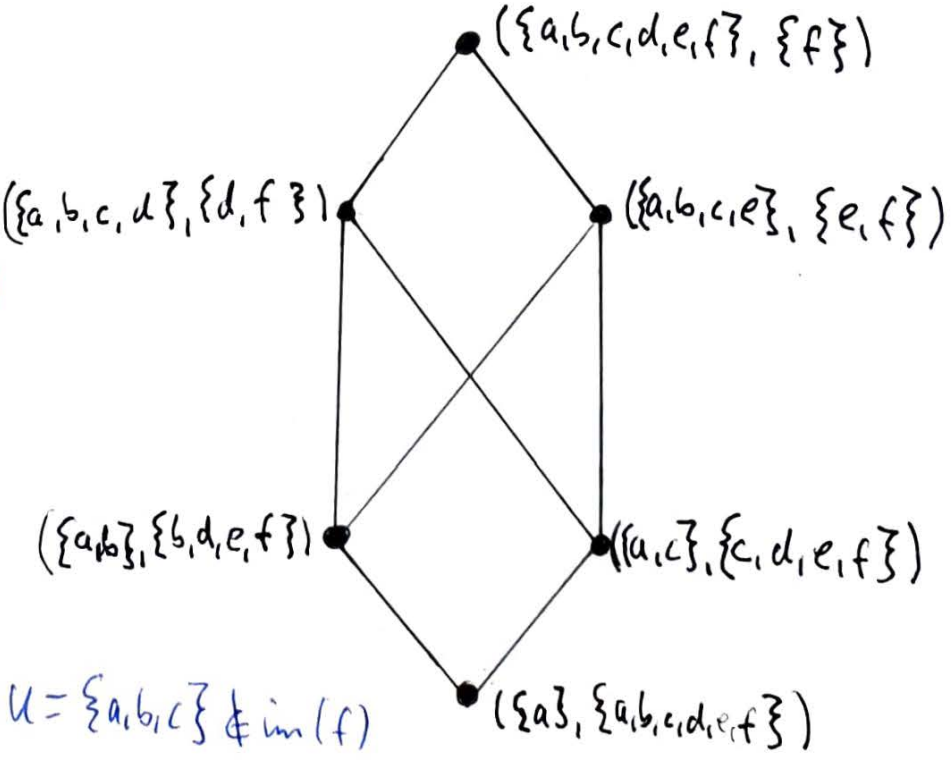
$(\downarrow x, \uparrow x) \sqsubseteq (\downarrow x', \uparrow x')$

$:\Leftrightarrow \downarrow x \subseteq \downarrow x' \ (\Leftrightarrow \uparrow x \supseteq \uparrow x')$

$O = \{d,e,f\} \supsetneq \tilde{O}$



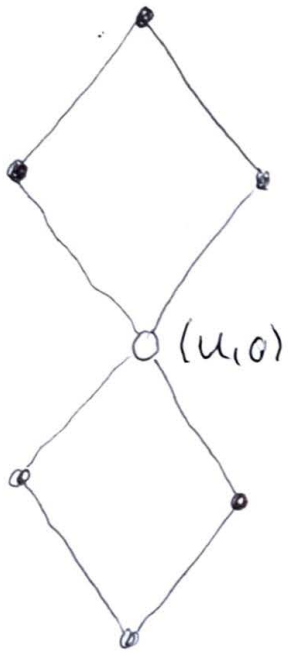
$\Rightarrow U = \{a,b,c\} \notin \mathrm{im}(f)$



$\Rightarrow$ zusätzliches „ideales Element" $(\{a,b,c\}, \{d,e,f\})$

- Das ist das einzige „ideale" Element, das noch zu $\text{im}(g)$ hinzukommt

$\Rightarrow$ $\text{im}(g) \cup \{(U,0)\}$ mit $(U,0) = (\{a,b,c\}, \{d,e,f\})$ sieht so aus:

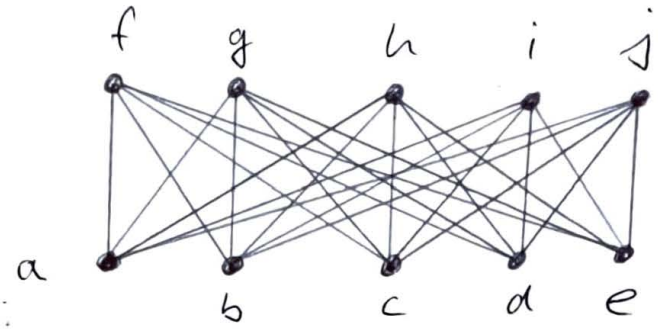


- Das sieht schon etwas aufgeräumter aus

als 

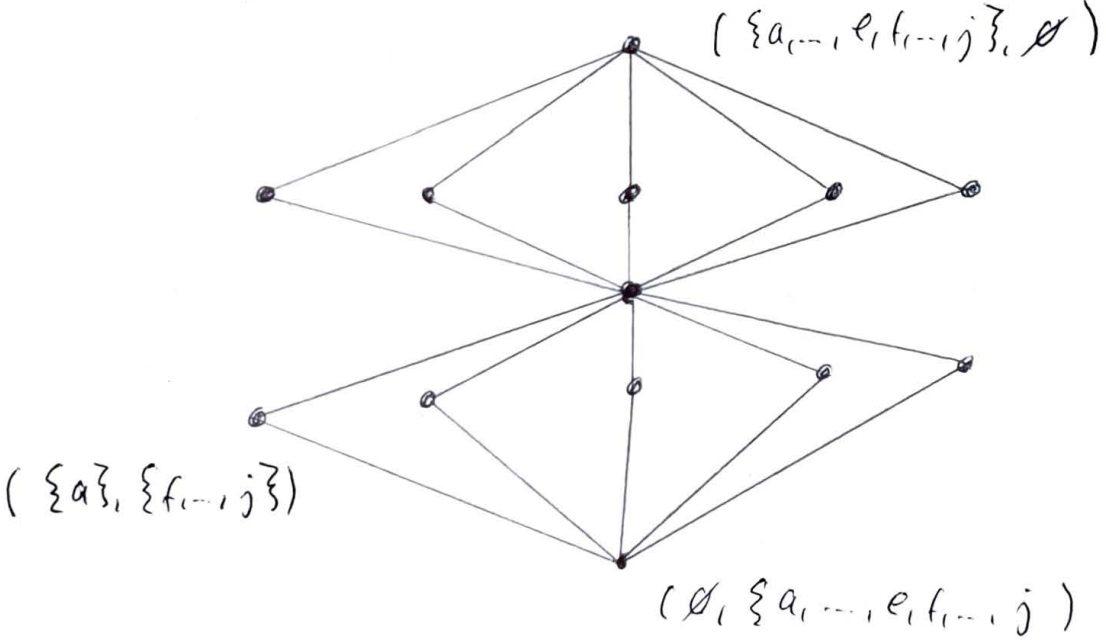
Anderes Beispiel

$(X, \leq)$:

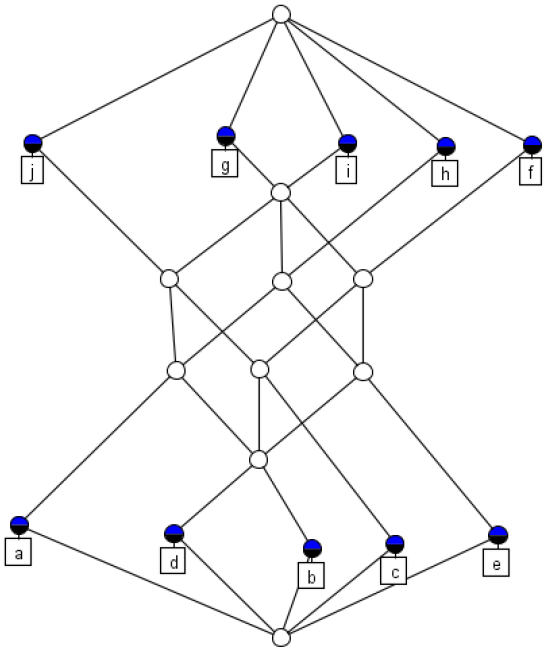


($x \leq y$ für jedes $x \in \{a, \ldots, e\}$ und jedes $y \in \{f, \ldots, j\}$.)

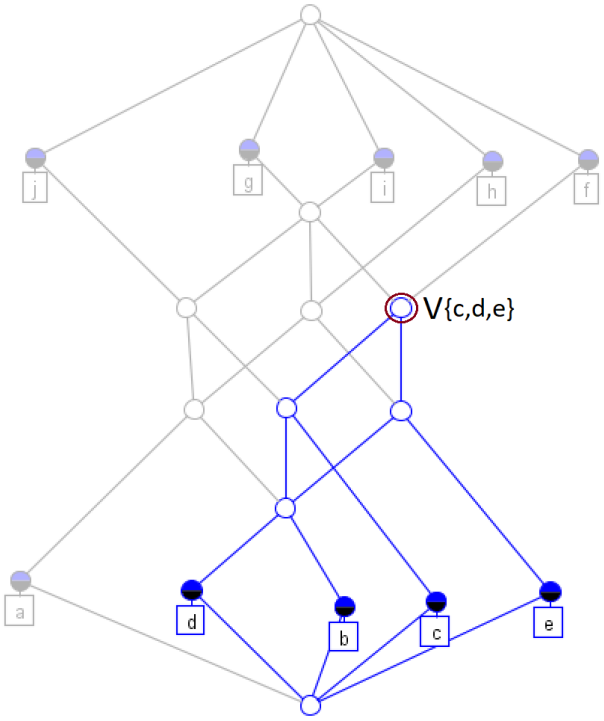
$\Rightarrow$ Menge der idealen Elemente:



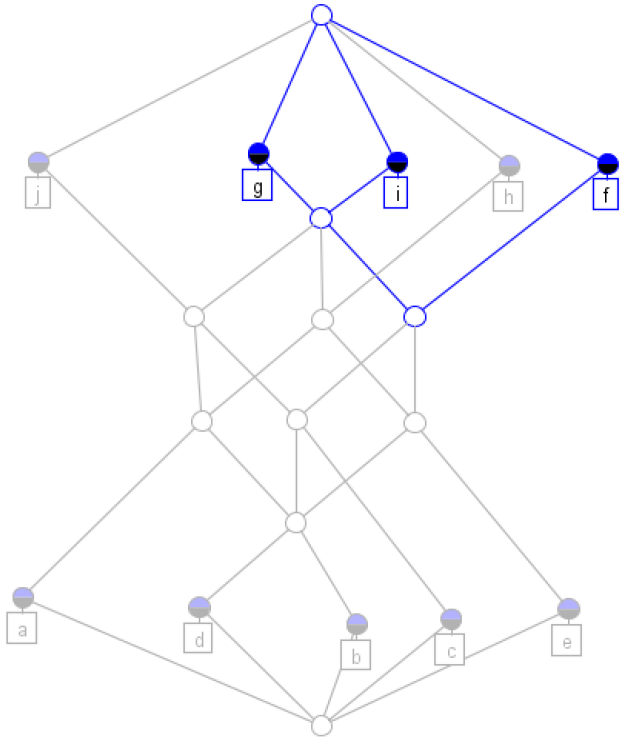
... Sieht auch aufgeräumter aus. Wenn man z.B. die Paare $(a, f), (c, h)$ und (e, j) aus $(X, \leq)$ entfernt, dann wäre es im Hassediagramm von $(X, \leq)$ sicherlich schwer, dies graphisch gut zu erkennen. Die idealen Elemente der so veränderten geordneten Menge sehen jedoch immer noch recht übersichtlich aus:



Wenn man jetzt beispielsweise alle („realen“ wie „idealen“) oberen Schranken der Menge $S = \{c, d, e\}$ bestimmen will, dann kann man die kleinste obere Schranke von S suchen (diese existiert dann immer, hier als „ideales“ Element, vergleiche untenstehende Bemerkungen).



Die oberen Schranken von S sind dann genau alle Elemente oberhalb dieses Supremums, also die „realen“ Elemente f, g und i sowie zwei weitere „ideale“ Elemente:



Bemerkungen:

i) Die Menge aller „idealen“ Elemente bezeichnen wir mit $\mathfrak{B}((X, X, \leq))$.

ii) Für $x \in X$ ist (U_x, O_x) mit $U_x := \downarrow x$ und $O_x := \uparrow x$ ein „ideales“ Element.

iii) Für $x, x' \in X$ gilt

$$x \leq x' \iff \downarrow x \subseteq \downarrow x' \iff \uparrow x \supseteq \uparrow x'.$$

iv) Für beliebige „ideale“ Elemente (U, O) und (U', O') gilt

$$\begin{aligned} O \subseteq O' &\Longrightarrow \Phi(O) \supseteq \Phi(O') \\ &\iff U \supseteq U' \\ &\Longrightarrow \Psi(U) \subseteq \Psi(U') \\ &\iff O \subseteq O', \end{aligned}$$

also insgesamt: $O \subseteq O' \iff U \supseteq U'$.

v) $\mathfrak{B}((X, X, \leq))$, ausgestattet mit der Ordnungsrelation $\sqsubseteq$ gegeben durch

$$(U, O) \sqsubseteq (U', O') :\iff U \subseteq U' \iff O \supseteq O'$$

ist ein vollständiger Verband, in dem das Infimum gegeben ist durch

$$\bigsqcap_{i \in I} (U_i, O_i) = \left(\bigcap_{i \in I} U_i, \Psi \left(\bigcap_{i \in I} U_i \right) \right) = \left(\bigcap_{i \in I} U_i, (\Psi \circ \Phi) \left(\bigcup_{i \in I} O_i \right) \right).$$

Das Supremum ist gegeben durch

$$\bigsqcup_{i\in I}(U_i,O_i)=\left(\Phi\left(\bigcap_{i\in I}O_i\right),\bigcap_{i\in I}O_i\right)=\left((\Phi\circ\Psi)\left(\bigcup_{i\in I}U_i\right),\bigcap_{i\in I}O_i\right).$$

(Begründung: z.B. für das Infimum gilt: $\bigcap\limits_{i\in I}U_i=\bigcap\limits_{i\in I}\left(\bigcap\limits_{o\in\tilde{O}_i}\downarrow o\right)=\bigcap\limits_{o\in\bigcup\limits_{i\in I}\tilde{O}_i}\downarrow o$ ist ein zulässiges „Unterhalb“ und nach Konstruktion das größte „Unterhalb“, das Teilmenge von jedem U_i ist.

Satz 1.17 (Dedekind-Mac-Neille-Vervollständigung)
Jede geordnete Menge $(X,\leq)$ lässt sich in den vollständigen Verband $\mathfrak{B}((X,X,\leq))$ vermöge der Abbildung

$$f:(X,\leq)\longrightarrow\mathfrak{B}((X,X,\leq)):x\mapsto(\downarrow x,\uparrow x)$$

einbetten. Dabei gilt insbesondere, dass f existierende Suprema und Infima in $(X,\leq)$ erhält.

Bemerkung 1.16. Die Dedekind-Mac-Neille-Vervollständigung ist in gewissem Sinne die sparsamste Einbettung, genauer gesagt existiert für jede weitere Einbettung $g:(X,\leq)\longrightarrow(Z,\sqsubseteq)$ in einen vollständigen Verband $(Z,\sqsubseteq)$ eine Ordnungseinbettung $\lambda:\mathfrak{B}((X,X,\leq))\longrightarrow(Z,\sqsubseteq)$ mit $g=\lambda\circ f$.

Wichtige Bemerkung: Auch hier ist die Tatsache, dass wir eine homogene Ordnungsrelation betrachteten, ohne Belang! Im Folgenden werden wir nun allgemein inhomogene Inzidenzstrukturen (G,M,I) betrachten. Wir werden genau die gleiche Konstruktion von „idealen“ Elementen betrachten und so zum sogenannten Begriffsverband eines formalen Kontetxtes gelangen, wir beginnen also jetzt mit der Formalen Begriffsanalyse:

2 Formale Begriffsanalyse

Die formale Begriffsanalyse ist eine von der Darmstädter Forschungsgruppe um Rudolf Wille, Bernhard Ganter und Peter Burmeister Anfang der 1980er Jahre entwickelte mathematische Theorie, die sich als angewandte Ordnungs- und Verbandstheorie versteht. Intellektuell geprägt wurde die Entwicklung der Formalen Begriffsanalyse durch die Auffassungen von Hartmut von Hentig und seiner Forderung nach einer Restrukturierung der Wissenschaften:

„... dann müssen die einzelnen Wissenschaften in erster Linie ihre Disziplinarität überprüfen, und das heißt, ihre unbewußten Zwecke aufdecken, ihre bewußten Zwecke deklarieren, ihre Mittel danach auswählen und ausrichten und ihre Berechtigung, ihre Ansprüche, ihre möglichen Folgen öffentlich und verständlich darlegen und dazu ihren Erkenntnisweg und ihre Ergebnisse über die Gemeinsprache (und die von mir sogenannte 'Anschauung') zugänglich machen.“ (von Hentig 1974, S. 136)

Im Sinne dieser Forderung argumentiert auch Rudolf Wille im Abstract seines Aufsatzes *Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts*:

„Lattice theory today reflects the general status of current mathematics: there is a rich production of theoretical concepts, results, and developments, many of which are reached by elaborate mental gymnastics; on the other hand, the connections of the theory to its surroundings are getting weaker and weaker, with the result that the theory and even many of its parts become more isolated. Restructuring lattice theory is an attempt to reinvigorate connections with our general culture by interpreting the theory as concretely as possible, and in this way to promote better communication between lattice theorists and potential users of lattice theory.“

Die rein mathematischen Grundlagen der formalen Begriffsanalyse wurden bereits in den 1930er Jahren von Garrett Birkhoff im Rahmen einer allgemeinen Verbandstheorie gelegt. Ausgangspunkt der Formalen Begriffsanalyse ist die mathematische Formalisierung des Begriffs „Begriff“, die unter anderem auch von den Schriften von Charles Sanders Peirce inspiriert ist (Deutsche Übersetzung: Klaus Oehler 1968):

„ 389. Ein klarer Begriff wird definiert als ein solcher, der so erfaßt ist, daß er wiedererkannt wird, wo er auch angetroffen werden mag, und der so erfasst ist, daß kein anderer Begriff mit ihm verwechselt wird. Wenn er dieser Klarheit ermangelt, nennt man ihn dunkel ...“

„... Andererseits, bloß derart eine Kenntnis des Begriffs zu haben, daß man mit ihm vertraut geworden ist und alles Zögern, ihn in gewöhnlichen Fällen wiederzuerkennen, verloren hat, scheint kaum den Namen der Klarheit des Begreifens zu verdienen, da es schließlich nur auf ein subjektives Gefühl der Beherrschung hinausläuft, welches völlig im Irrtum sein kann.“

„390. Ein deutlicher Begriff wird definiert als ein solcher, der nichts enthält, was nicht klar ist. Das ist technische Sprache. Unter dem Inhalt eines Begriffs verstehen die Logiker alles, was in seiner Definition enthalten ist. So daß gemäß ihrer Auffassung ein Begriff dann deutlich erfasst ist, wenn wir in abstrakten Termini eine präzise Definition von ihm geben können.

An diesem Punkt lassen die Berufslogiker die Sache auf sich beruhen, und ich hätte den Leser nicht mit dem belästigt, was sie zu sagen haben, wenn das nicht ein so schlagendes Beispiel dafür wäre, wie sie ganze Zeitalter geistiger Tätigkeit verschlafen haben, die Maschinerie des modernen Denkens gleichgültig übersahen und sich nicht im Traum einfallen ließen, sie zur Verbesserung der Logik anzuwenden. Man kann leicht zeigen, daß die Lehre, daß vertrauter Gebrauch und abstrakte Deutlichkeit die Vollkommenheit des Begreifens ausmachen, ihren einzigen richtigen Platz in Philosophien hat, welche schon lange tot sind. Es ist jetzt an der Zeit, die Methode zu formulieren, mit der man eine größere Klarheit des Denkens erreichen kann, so wie wir sie bei den Denkern unserer Zeit sehen und bewundern ...“

„... Nichts Neues kann je durch die Analyse von Definitionen erkannt werden. Trotzdem können unsere bestehenden Überzeugungen durch dieses Verfahren geordnet werden, und Ordnung ist ein wesentliches Element der Ökonomie des Denkens, wie jeder anderen auch. Es mag deshalb anerkannt werden, daß die Lehrbücher im Recht sind, wenn sie die Vertrautheit mit einem Begriff zu dem ersten Schritt in Richtung auf die Klarheit des Erfassens machen und seine Definition zu dem zweiten. Aber dadurch, daß sie jede Erwähnung irgendeines höheren Grades der Klarheit des Denkens unterlassen, spiegeln sie einfach eine Philosophie wieder, welche schon vor hundert Jahren zertrümmert worden ist. Jene vielbewunderte 'Zierde der Logik' - die Lehre von der Klarheit und Deutlichkeit - mag ganz hübsch sein, aber es ist höchste Zeit, den antiken Edelstein in unser Kuriositätenkabinett zu verbannen und etwas an uns zu tragen, das für modernen Gebrauch besser geeignet ist. ...“

„402. Es scheint mithin, daß die Regel zur Erlangung des dritten Grades der Klarheit des Erfassens folgende ist: Überlege, welche Wirkungen, die denkbarerweise praktische Bezüge haben könnten, wir dem Gegenstand unseres Begriffs in Gedanken zukommen lassen. Dann ist unser Begriff dieser Wirkungen das Ganze unseres Begriffs des Gegenstandes.“ [Peirce, 1878]

Die ersten beiden Forderungen nach der Klarheit und Deutlichkeit von Begriffen finden unmittelbar in der mathematischen Definition eines formalen Begriffes Niederschlag, vom dritten Grade der Klarheit des Erfassens, von der sogenannten Pragmatischen Maxime, wird an anderer Stelle unter anderem behauptet[4]:

„In that tradition, FCA aims at unfolding the observable, elementary properties defining the objects subsumed by scientific concepts.“ [Wollbold, 2012]

[4]Was hier genau mit *unfolding* gemeint ist, und inwieweit man wirklich wenigstens in die Nähe des dritten Grades der Klarheit des Erfassens gelangt, ist jedenfalls mir nicht so ganz ersichtlich.

Definition 2.1 (Formaler Kontext)
Ein **formaler Kontext** ist ein Tripel (G, M, I), wobei G und M (nichtleere) Mengen und $I \subseteq G \times M$ eine binäre Relation zwischen G und M sind. Ein formaler Kontext ist also nichts anderes als eine Inzidenzstruktur. Die Menge G kann man sich als Menge von Gegenständen und die Menge M kann man sich als Menge von Merkmalen vorstellen. Die Aussage gIm ist dann zu interpretieren als: „Gegenstand g besitzt Merkmal m“.

Definition 2.2 (Formaler Begriff)
Ein **formaler Begriff** ist ein Paar (A, B), wobei A eine Menge von Gegenständen und B eine Menge von Eigenschaften ist, so dass gilt:

i) Jeder Gegenstand g aus A besitzt jedes Merkmal m aus B, in Zeichen: $\forall g \in A, \forall m \in B : gIm$.

ii) Es gibt kein weiteres Merkmal $m \in M \backslash B$, dass alle Gegengstände g aus A gemeinsam besitzen, d.h. formal: $\forall m \in M : (\forall g \in A : gIm) \Longrightarrow m \in B$.

iii) Es gibt keinen weiteren Gegenstand $g \in G \backslash A$, der ebenfalls alle Merkmale aus B besitzt, also: $\forall g \in G : (\forall m \in B : gIm) \Longrightarrow g \in A$.

Bemerkung 2.1. Wenn man sich den formalen Kontext (G, M, I) als Kreuzchentabelle vorstellt, dann sind die formalen Begriffe als „maximale, mit Kreuzchen gefüllte Rechtecke“ vorstellbar. Die Menge A von Gegenständen wird auch als **Begriffsumfang** bzw. **Begriffsextension** und die Menge B von Merkmalen wird auch als **Begriffsinhalt** bzw. **Begriffsintension** bezeichnet.

Es folgen nun ein paar Überlegungen, die wir eigentlich bereits im Abschnitt zu den „idealen“ Elementen im Kontext der Ordnungsrelationen angestellt haben, das Folgende ist also im Wesentlichen nichts Neues:

Beobachtungen:

- Der Schnitt zweier Begriffsumfänge ist wieder ein Begriffsumfang.
- Der Schnitt zweier Begriffsinhalte ist wieder ein Begriffsinhalt.
- Aber: Der Schnitt zweier maximaler Rechtecke ist im Allgemeinen kein maximales Rechteck mehr.
- Außerdem: Die Vereinigung von zwei Begriffsumfängen (Begriffsinhalten) ist nicht immer ein Begriffsumfang (Begriffsinhalt).

Fragen:

- Warum ist es sinnvoll, nur maximale Rechtecke zu betrachten?
- Wie kann man eine „schöne“ Ordnung auf der Menge der formalen Begriffe definieren bzw.
- Bildet die noch zu definierende Ordnungsrelation dann eine (vollständige) Verbandsstruktur?
- Wir wollen eine Ordnung einführen, die modellieren soll, dass ein Begriff (A, B) ein Unterbegriff eines anderen Begriffes (C, D) ist. Dazu sollte sicherlich der Unterbegriff spezifischer sein, d.h., er sollte wohl weniger Gegenstände umfassen, die dafür dann aber mehr gemeinsame Merkmale besitzen. Eine denkbare Definition einer Unterbegriffsrelation ist also

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

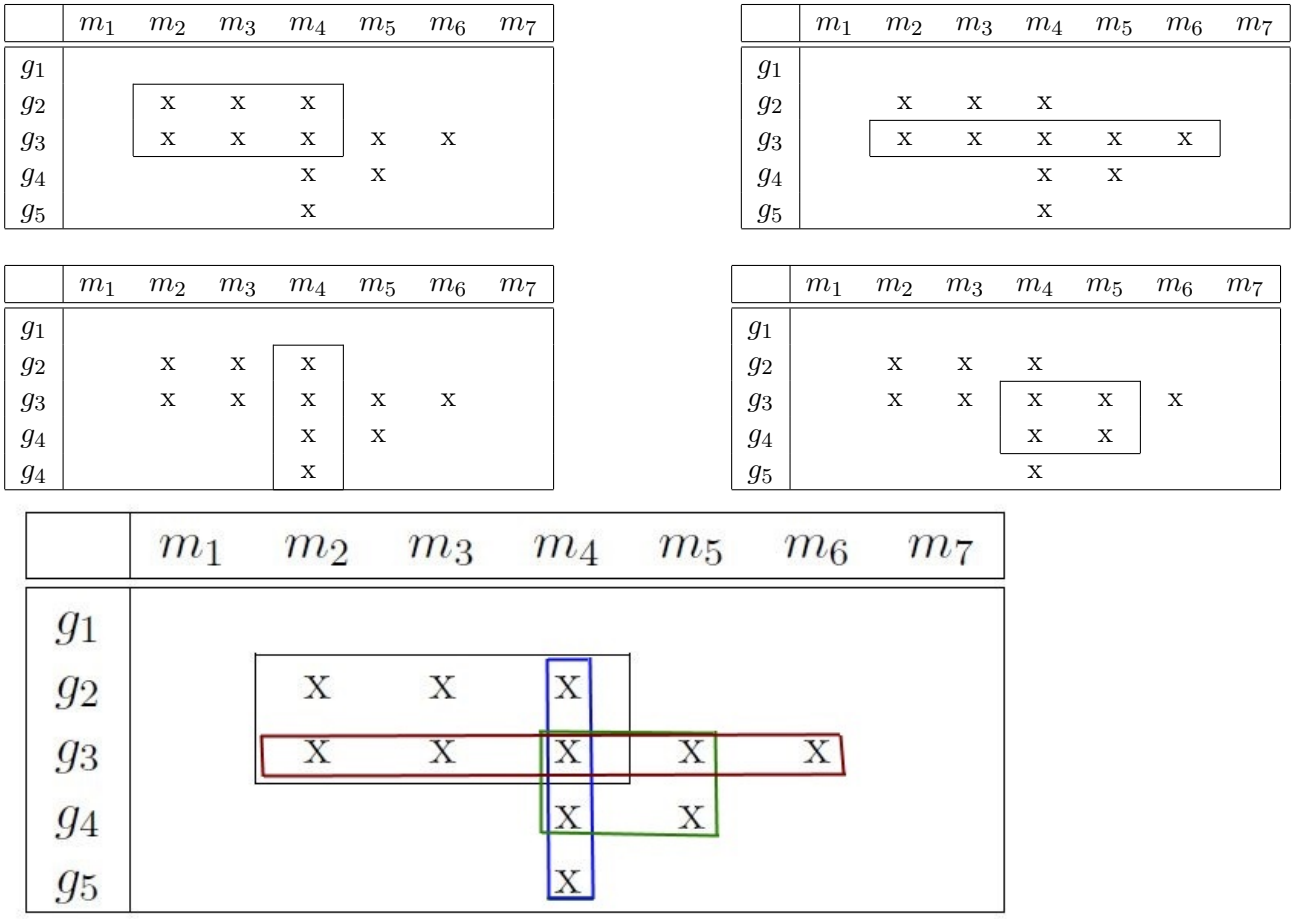


Abbildung 2: Beispiel eines formalen Kontextes dargstellt als Kreuzchentabelle mit 4 formalen Begriffen dargestellt als maximale Rechtecke. (Insgesamt gibt es hier jedoch 6 formale Begriffe, welche fehlen noch?)

- Dann stellt sich die Frage, ob diese Definition einen vollständigen Verband von Begriffen definiert.
- Da der Schnitt von Begriffsumfängen wieder ein Begriffsumfang ist, liegt die Vermutung nahe, dass zwei Begriffe einen größten gemeinsamen Unterbegriff besitzen, dessen Umfang durch den Schnitt der Umfänge dieser beiden Begriffe gegeben ist.
- Wenn wir den Umfang eines beliebigen Begriffs kennen, dann können wir ganz einfach den Inhalt des Begriffs als Menge aller Merkmale, die alle Gegenstände des Umfangs gemeinsam besitzen, ableiten. Dies tut genau die Ableitungsoperation Ψ.
- Andersherum können wir bei bekanntem Begriffsinhalt den zugehörigen Begriffsumfang als Menge aller Gegenstände, die alle Merkmale des Begriffsinhalts besitzen, ableiten. Dies tut genau die Ableitungsoperation Φ.
- Für den kleinsten gemeinsamen Obergbegriff zweier Begriffe kann man zunächst nicht direkt mit der Vereinigung arbeiten, da die Vereinigung von Umfängen im Allgemeinen kein Umfang mehr ist, man kann aber, anstelle die Umfänge zu vereinen, einfach die Inhalte schneiden, um den Begriffsinhalt des kleinsten gemeinsamen Oberbegriffs zu erhalten. Dann kann man anschließend mit dem Operator Φ analog den Begriffsumfang des kleinsten gemeinsamen Oberbegriffs erhalten.

Der folgende Hauptsatz der formalen Begriffsanalyse präzisiert und vervollständigt obenstehende Gedanken:

Satz 2.3 (Hauptsatz der formalen Begriffsanalyse)
Sei (G, M, I) ein formaler Kontext und $\mathfrak{B}(G, M, I)$ die Menge aller formalen Begiffe, die mit der Ordnungsrelation $\leq$, gegeben durch

$$(A, B) \leq (C, D) : \Longleftrightarrow \ A \subseteq C \ \& \ B \supseteq D$$

ausgestattet sei. Seien weiter die Operatoren

$$\Phi : (2^M, \subseteq) \longrightarrow (2^G, \supseteq) : \quad B \mapsto B' := \underbrace{\{g \in G \mid \forall m \in B : gIm\}}_{\substack{\text{Menge aller Gegenstände, die alle} \\ \text{Merkmale aus } B \text{ besitzen.}}} \text{ und}$$

$$\Psi : (2^G, \supseteq) \longrightarrow (2^M, \subseteq) : \quad A \mapsto A' := \underbrace{\{m \in M \mid \forall g \in A : gIm\}}_{\substack{\text{Menge aller den Gegenständen aus} \\ A \text{ gemeinsamer Merkmale.}}}.$$

im Folgenden jeweils immer abgekürzt mit $'$. (Ob es sich um Φ oder Ψ handelt, wird immer eindeutig aus dem Kontext heraus klar sein.) Dann bildet $\mathfrak{B}(G, M, I)$ einen vollständigen Verband, in dem Infimum und Supremum einer Klasse $(A_t, B_t)_{t \in T}$ von Begriffen gegeben sind durch

$$\bigwedge_{t \in T} (A_t, B_t) = \left(\bigcap_{t \in T} A_t \ , \ \left(\bigcap_{t \in T} A_t \right)' \right) = \left(\bigcap_{t \in T} A_t \ , \ \left(\bigcup_{t \in T} B_t \right)'' \right)$$

$$\bigvee_{t \in T} (A_t, B_t) = \left(\left(\bigcap_{t \in T} B_t \right)' \ , \ \bigcap_{t \in T} B_t \right) = \left(\left(\bigcup_{t \in T} A_t \right)'' \ , \ \bigcap_{t \in T} B_t \right).$$

Zur Berechnung des Begriffsverbandes Eine einfache, aber sehr aufwendige Möglichkeit: Bestimme für jede Menge $Y \subseteq M$ von Merkmalen die Menge Y' aller Gegenstände, die diese Menge von Merkmalen gemeinsam haben. Bestimme anschließend die Menge Y'' aller Merkmale, die diesen Gegenständen wiederum gemeinsam sind (, das können durchaus mehr als die ursprünglichen Merkmale aus Y sein). Also: Bestimme für alle $Y \in 2^M$ den von Y erzeugten Begriff

$$(Y', Y''), \tag{4}$$

womit man in der Tat alle formalen Begriffe erhält. Dies ist natürlich sehr aufwendig, da viele formale Begriffe mehrmals erzeugt werden. Es gibt sehr viele sehr viel effizientere Algorithmen, die alle formalen Begriffe erzeugen. Einer davon ist der next-closure Algorithmus, beschrieben in [Ganter, 2013, S. 84-89], der allgemein alle Hüllen eines beliebigen Hüllenoperators berechnen kann.

2.1 Eine kleine Anwendung zur Illustration

„Seit ein Gespräch wir sind und hören voneinander,“ [Hölderlin]

Das folgende Beispiel soll kurz illustrieren, wie die Betrachtung von Begriffsverbänden für eine (rein deskriptive) Datenanalyse eingesetzt werden könnte. Das Beispiel ist entnommen aus [Ganter, 2013, S. 61ff]. Es geht dabei um die Krankheit Anorexia nervosa (Magersucht). Magersucht ist eine Form der Essstörung, bei der die Betroffenen eine veränderte Wahrnehmung des eigenen Körpers haben und aus Furcht vor Gewichtszunahme die Nahrungsaufnahme teilweise verweigern. Die Ursachen von Magersucht sind derzeit nicht vollständig geklärt, es scheint aber, dass Magersuchtpatienten oft eine veränderte Wahrnehmung ihres persönlichen sozialen Umfeldes haben. Eine Form der Therapie besteht daher darin, im Rahmen einer Gesprächstherapie die Sicht der Patienten auf ihr Umfeld in Therapiesitzungen zu beeinflussen. Dies ist natürlich sehr aufwendig, oft bedarf es vieler Sitzungen bis sich, wenn überhaupt, ein erster Erfolg einstellt. Um den Verlauf der Interviewgespräche in einer praktikablen, dem klinischen Alltag angemessenen Weise zu dokumentieren, werden unter anderem sogenannte Repertory Grids eingesetzt, bei der die Patienten nach einer standardisierten Methode nach ihren Einschätzungen zu ihrem Umfeld bezüglich gewisser Eigenschaften befragt werden. Tabelle 1 zeigt beispielhaft ein solches Repertory Grid (in leicht vereinfachter Darstellung):

	verletzlich	verschlossen	selbstsicher	pflichtbewusst	herzlich	schwierig	aufmerksam	leicht beleidigt	nicht jähzornig	ängstlich	geschwätzig	oberflächlich	sensibel	ehrgeizig
Ich selbst	x	x	x		x	x	x		x	x			x	x
Mein Ideal	x		x	x	x		x		x				x	x
Vater	x	x		x	x	x	x	x	x	x		x	x	x
Mutter	x	x		x	x	x		x	x	x		x	x	x
Schwester	x	x		x	x	x	x		x	x			x	x
Schwager			x	x	x		x				x	x		x

Tabelle 1: Repertory Grid, entnommen aus [Ganter, 2013, S. 62]

Beispielsweise das Kreuz in Zeile 1 und Spalte 1 würde bedeuten, dass die betroffene Person sich selbst als verletzlich einschätzen würde. (Natürlich könnte man diese Einschätzung nicht nur auf einem binären, sondern beispielsweise auch auf einem ordinalen Skalenniveau abfragen, was durchaus gemacht wird. Dazu allgemein mehr beim späteren Abschnitt zum sogenannten begrifflichen Skalieren.)

Nun hätte man zu solchen Repertory Grids auch noch gern anschauliche Graphiken, die die aktuelle Situation eines Patienten schnell und übersichtlich zusammenfassen. Dazu benutzt man sogenannte Biplots. Zur Illustration dieser graphischen Methode sei ein kleines fiktives (zugegebenermaßen wohl etwas ins Polemische geratene) Gespräch zwischen Arzt und Patient gestattet:

Therapeut: *„Ich sehe hier in meinen Dokumenten, dass, ... , es scheint mir so, als ob Sie sich in den letzten Wochen wieder mehr und mehr Ihrer Schwester 'nahe' fühlen.“*

Patient: *„Das ist mir noch nicht so recht aufgefallen und irgendwie fühlt es sich auch überhaupt nicht so an, ich und meine Schwester hatten eigentlich immer die gleiche starke Bindung zueinander. Woraus schließen Sie Ihre Beobachtung denn?“*

Therapeut: *„Wir nutzen da diese graphische Methode, dass nennt sich Biplot, ist vielleicht etwas kompliziert zu erklären, die grobe Idee ist einfach die, dass wir Ihre Einschätzungen zu Ihrem sozialen Umfeld in einem hochdimensionalen Raum darstellen und uns dann die interessantesten Dimensionen dieses Raumes anschauen. Mathematisch schaut man einfach aus einer geschickt gewählten Perspektive auf eine hochdimensionale Datenwolke.“*

Patient: *„Was sind denn diese ominösen Dimensionen? Ein bisschen was von Mathematik verstehe ich schon, hab mal ein Semeseter lang Maschinenbau studiert.“*

Therapeut: *„Das sind einfach Linearkombinationen Ihrer Einschätzungen zu den jeweiligen Attributen.“*

Patient: *„Okay, das klingt sehr interessant aber irgendwie auch befremdlich. Sie sagten ich würde mich näher bei meiner Schwester verorten, ich nehme an, dass Sie da von einem Abstand in diesem hochdimensionalen Raum sprechen.“*

Therapeut: *„Genau richtig.“*

Patient: *„Ich hab mal was von Nichteuklidischer Geometrie und dem Begriff des euklidischen Abstands gehört, scheint als ob es auch irgendwie nichteuklidische Abstände gibt, ich meine mich zu erinnern, dass ein damaliger Studienkollege mal was von der Taximetrik erzählt hat, ich kann mich nicht mehr genau erinnern, aber irgendwie betrachtet man da nicht direkt den Abstand per Luftlinie, sondern den kürzesten Weg entlang eines rechteckigen Straßennetzes.... Da hat man wohl viele Möglichkeiten, Ihre Dokumente auf graphische Weise zu studieren.“*

Therapeut: *„Ja, man hat hier viel Flexibilität.“*

Patient: *„Klingt alles auch irgendwie willkürlich, naja, solange man diese Biplots schön inhaltlich interpretieren kann, dann ist das sicherlich interessant. Wie würde man Ihre Aussage über mich und meine Schwester denn ganu inhaltlich interpretieren?“*

Therapeut: *„Ganz einfach: Der Abstand ist der Abstand ist der Abstand, eben in diesem hochdimensionalen Raum. Und dieser ist eben scheinbar kleiner geworden.“*

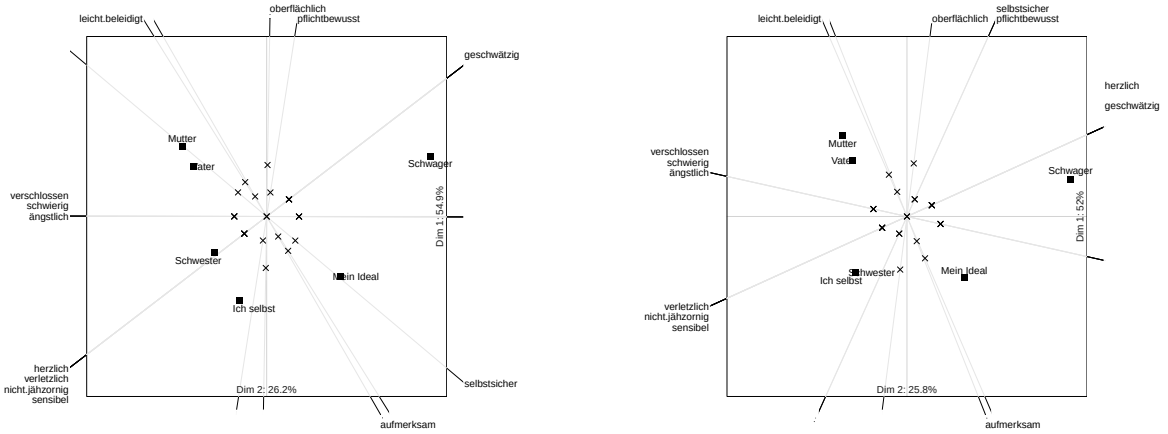


Abbildung 3: Biplots zu den Repertory Grids zur Situation von vor ein paar Wochen (links) und zur aktuellen Situaton (rechts).

Patient: *„Aber was bedeutet es jetzt genau, dass ich mich näher bei meiner Schwester sehe?“*

Therapeut: *„Okay, dann müsste ich wohl doch noch einmal direkt in die originalen Interview-Daten schauen, der Biplot scheint Ihre Frage dann doch nicht so recht beantworten zu wollen.“*

Patient: „Und warum arbeiten Sie dann nicht gleich direkt mit den Original-Daten?“

Therapeut: *„Weil die sehr umständlich zu lesen sind, der Mensch ist halt ein Augentier.“*

... Ein Blick in die Repertory Grids hätte den Herrn Therapeuten schnell davon überzeugt, dass sich an der Selbstbeurteilung seines Gegenübers in den letzten Wochen ebenso nichts gändert hatte wie an der Einschätzung der Schwester... Lediglich die ach so geschickt gewählte Perspektive hatte sich einigermaßen unbemerkt verschoben:

	verletzlich	verschlossen	selbstsicher	pflichtbewusst	herzlich	schwierig	aufmerksam	leicht beleidigt	nicht jähzornig	ängstlich	geschwätzig	oberflächlich	sensibel	ehrgeizig
Ich selbst	x	x	x		x	x	x		x	x			x	x
Mein Ideal	x		x	x	x		x		x				x	x
Vater	x	x		x	x	x	x	x	x	x		x	x	x
Mutter	x	x		x	x	x		x	x	x		x	x	x
Schwester	x	x		x	x	x	x		x	x			x	x
Schwager			x	x	x		x				x	x		x

Tabelle 2: Repertory Grid von vor ein paar Wochen.

	verletzlich	verschlossen	selbstsicher	pflichtbewusst	herzlich	schwierig	aufmerksam	leicht beleidigt	nicht jähzornig	ängstlich	geschwätzig	oberflächlich	sensibel	ehrgeizig
Ich selbst	x	x	x		x	x	x		x	x			x	x
Mein Ideal	x		x	x	x		x		x				x	x
Vater	x	x	x	x	x	x	x	x	x	x		x	x	x
Mutter	x	x	x	x	x	x		x	x	x		x	x	x
Schwester	x	x		x	x	x	x		x	x			x	x
Schwager			x	x	x		x				x	x		x

Tabelle 3: Aktuelles Repertory Grid: Lediglich die Einschätzung von Vater und Mutter als nun selbstsicher hatte sich geändert.

Obiges Geschichtchen möchte verdeutlichen, dass Methoden wie Biplots allgemein schwer zu lesen bzw. im Zweifelsfall sogar irreführend sein können, nicht jedes „Bild von den Dingen hat eine natürliche Entsprechung in der Wirklichkeit der Daten“. Außerdem kann man aus den Biplots selbst die Originaldaten nicht wiedergewinnen. Schauen wir nun einmal, wie man die Daten mit Methoden der formalen Begriffsanalyse beschauen könnte. Dazu betrachten wir einfach obiges Repertory Grid als formalen Kontext und schauen uns den zugeörgen Begriffsverband an:

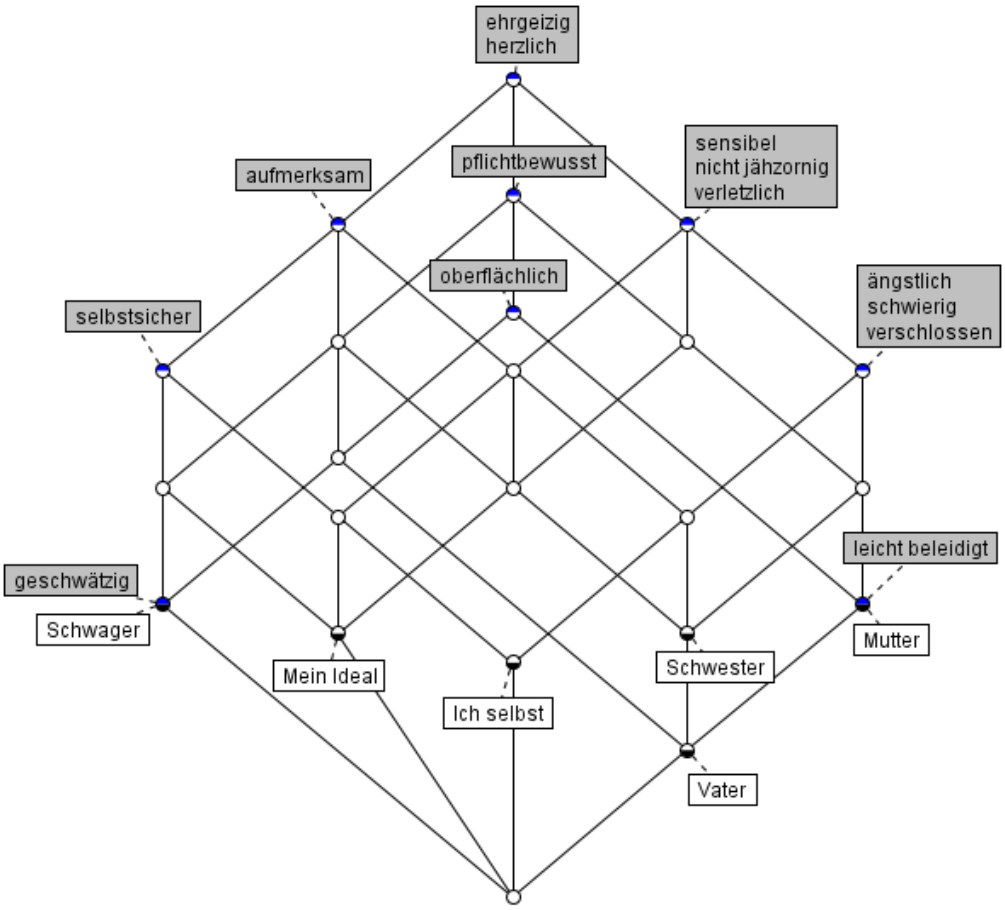


Abbildung 4: Vollständiger Verband aller formalen Begriffe des formalen Kontetxs zum Anorexia nervosa Beispiel.

Dieses Bild ist wie folgt zu lesen: Jeder Punkt ist ein formaler Begriff. Der zugehörige Begriffsinhalt ist gegeben durch alle Attribute, die an allen möglichen Oberbegriffen des Begriffs angetragen sind. Beispielsweise der formale Begriff links unten besitzt den Begriffsinhalt {geschwätzig, selbstsicher, aufmerksam, ehrgeizig, herzlich, pflichtbewusst, oberflächlich}. Der Begriffsumfang ist ablesbar als die Menge aller Gegenstände, die an allen möglichen Unterbegriffen stehen, hier konkret ist das lediglich der Schwager. Jetzt kann man offensichtlich die Originaltabelle aus der Graphik wiedergewinnen. Beispielsweise besitzt der Schwager (gemäß der Einschätzung des Patienten) genau die Eigenschaften geschwätzig, selbstsicher, aufmerksam, ehrgeizig, herzlich, pflichtbewusst und oberflächlich. Außerdem kann man recht leicht ablesen, welche Eigenschaften beispielsweise „Mein Ideal“ und „Ich selbst“ gemeinsam besitzen, indem man einfach die Eigenschaften, die zum spezifischsten gemeinsamen Oberbegriff von „Mein Ideal“ und „Ich selbst“ gehören, abliest. Das wären in diesem Fall die Eigenschaften selbstsicher, aufmerksam, ehrgeizig, herzlich, sensibel, nicht jähzornig und verletzlich. Ebenso kann man auch die Gemeinsamkeiten von mehr als zwei Personen ablesen, indem man die entsprechenden Suprema im Begriffsverband anschaut.

Natürlich ist dies ein sehr einfaches Beispiel, oft sind die entsprechenden Graphiken rein aufgrund der schieren Anzahl von formalen Begriffen sehr viel schwieriger zu lesen, was sich bereits in der Betrachtung der veränderten aktuellen Situation in unserem Besipiel zeigt (siehe unten).

Es ist aber durchaus möglich, sich einen Begriffsverband nicht einfach nur über die Krücke einer graphischen Veranschaulichung (, ja, ja, der Mensch ist halt ein Augentier), sondern vor dem inneren (intellektuellen) Auge vorzustellen (dazu später mehr).

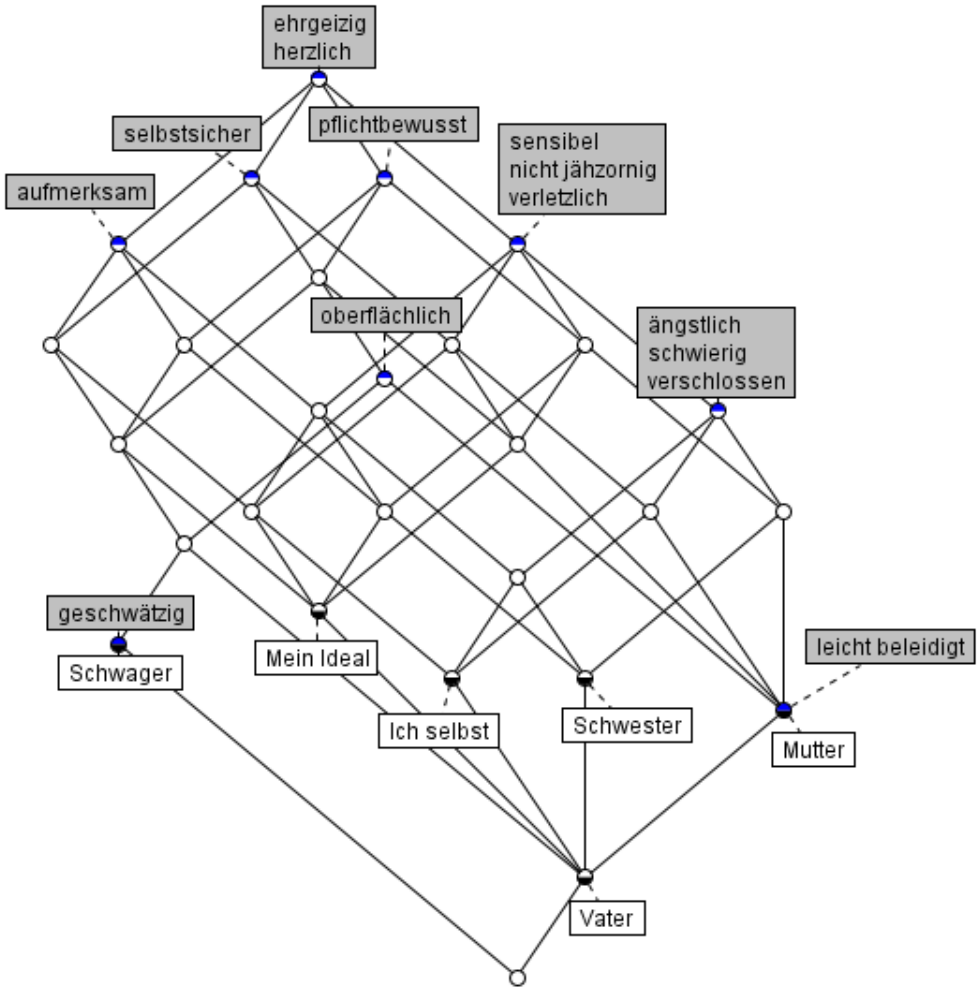


Abbildung 5: Vollständiger Verband aller formalen Begriffe des formalen Kontetxs zum Anorexia nervosa Beispiel für die aktuelle Situation.

2.2 Noch ein Beispiel: Punkte und Halbräume in $\mathbb{R}^2$

Sei der Kontext $\mathbb{K} := (G, M, I)$ gegeben durch

$G = \mathbb{R}^2 \ldots$ Menge aller Punkte in $\mathbb{R}^2$,
$M = \{\{x \in \mathbb{R}^2 \mid \langle x, v\rangle < c\} \mid v \in \mathbb{R}^2, c \in \mathbb{R}\} \ldots$ Menge aller offenen Halbräume[5] von $\mathbb{R}^2$,
$gIm \iff$ Punkt g liegt in Halbraum m.

Frage: Wie sehen dann die Umfänge des zugehörigen Begriffsverbands aus?

Antwort: Die Begriffsumfänge sind genau die konvexen Mengen von $\mathbb{R}^2$.

Grund:

i) Für jede konvexe Menge A gilt $A'' = A$, denn $A'' \supseteq A$ gilt immer, denn $''$ war ein Hüllenoperator und $A''\backslash A = \emptyset$ gilt wegen des Trennungssatzes für konvexe Mengen: Jede offene konvexe Menge A kann von jedem Punkt $g \notin A$ durch einen Halbraum getrennt werden in dem Sinne, dass ein $v \in \mathbb{R}^2$ und ein $c \in \mathbb{R}$ existieren, so dass $\langle x, v\rangle < c$ für alle $x \in A$, aber $\langle g, v\rangle > c$. (Wende diesen Satz auf das Innere einer beliebigen konvexen Menge an und betrachte den Halbraum $\{x \in \mathbb{R}^2 \mid \langle x, v\rangle < \tilde{c}\}$ mit $\tilde{c} := \langle g, v\rangle > c$. Dann liegt auch der Rand von A immer noch in diesem Halbraum, denn für Randpunkte x gilt immer noch $\langle x, v\rangle \leq c < \tilde{c}$ und g liegt trotzdem noch nicht in diesem Halbraum, denn $\langle g, v\rangle = \tilde{c} \not< \tilde{c}$.)

ii) Jeder Halbraum, der eine Menge A von Punkten enthält, enthält auch beliebige Konvexkombinationen von Punkten aus A, also die gesamte konvexe Hülle von A.

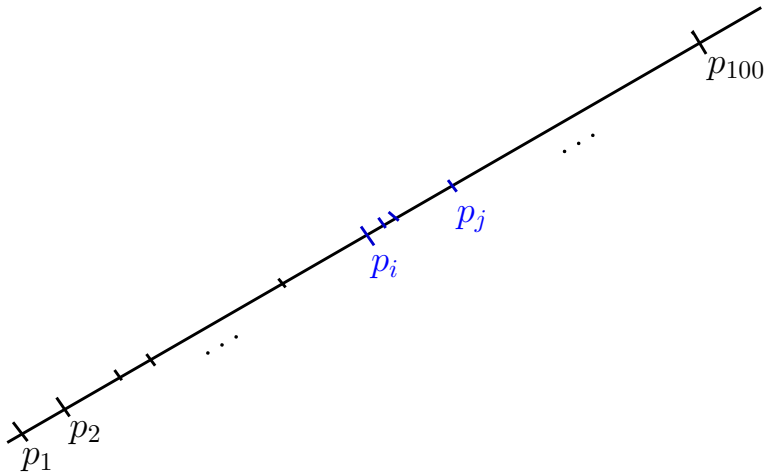
Wir werden jetzt obigen Kontext modifizieren, indem wir nicht alle Punkte von $\mathbb{R}^2$, sondern nur endlich viele, beispielsweise $n = 100$ Punkte von $\mathbb{R}^2$ als Gegenstände betrachten, es ist also $G = \{g_1, \ldots, g_{100}\}$. Dies ist insbesondere im statistischen Kontext interessant, wo man oft nur eine endliche Stichprobe von statistischen Einheiten (bzw. auch eine endliche Vollerhebung) betrachtet. Als Merkmale betrachten wir trotzdem weiterhin alle beliebigen offenen Halbräume von $\mathbb{R}^2$. (Man könnte hier auch mit endlich vielen geeignet gewählten Halbräumen arbeiten, namentlich mit denjenigen abgeschlossenen Halbräumen, die durch Geraden, die durch zwei Punkte g_i, g_j aus G verlaufen, beschrieben sind.) Dann sind die Begriffsumfänge nichts anderes als die auf die neue Gegenstandsmenge $G = \{g_1, \ldots, g_{100}\}$ projizierten konvexen Mengen: Jeder Umfang kann als

$$A \cap \{g_1, \ldots, g_{100}\}$$

mit einer geeigneten konvexen Menge A dargestellt werden. Je nachdem, wie die Punkte genau in der Ebene liegen, kann der entstehende Begriffsverband sehr verschieden groß sein, wir betrachten zwei extreme Beispiele:

i) Alle Punkte $g_1, \ldots, g_{100}$ liegen auf einer Geraden:

[5] Man könnte hier auch noch die abgeschlossenen Halbräume hinzunehmen bzw. auch beliebige konvexe Mengen betrachten, der Satz zur Trennung konvexer Mengen hat jedoch zur Folge, dass dies nichts ändern würde, die Menge aller offenen Halbräume reicht also schon völlig aus.

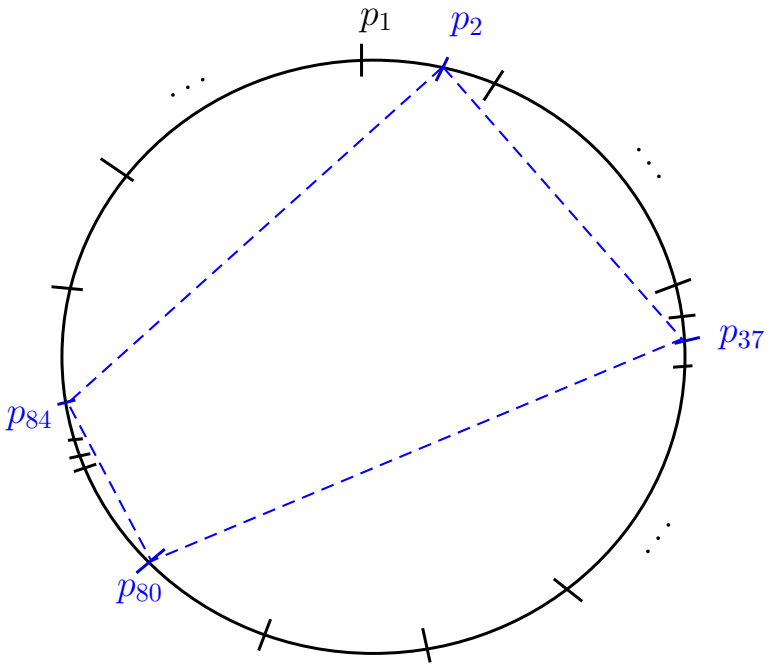


Dann ist die Menge $\mathfrak{BU}$ der Begriffsumfänge gegeben durch:

$$\mathfrak{BU} = \{\{g_i, g_{i+1}, \ldots, g_j\} \mid i, j \in \{1, \ldots, 100\}, i \leq j\} \cup \emptyset.$$

Hier gibt es also $\frac{100 \cdot 99}{2} + 100 + 1 = 5051$ Begriffe. Allgemein für n Punkte gäbe es $\frac{n(n-1)}{2} + n + 1$ Begriffe.

ii) Alle Punkte liegen auf einem Kreis, einer Ellipse oder anderweitig derart, dass kein Punkt in der konvexen Hülle von anderen Punkten liegt:



Dann ist offensichtlich nach obigen Überlegungen jede beliebige Teilmenge von Punkten ein formaler Begriffsumfang, es gibt also $2^{100} \approx 1.27 \cdot 10^{30}$, bzw. allgemein 2^n Begriffe, was sehr viel mehr ist. Zur Illustration wollen wir einen kleinen Datensatz im Zusammenhang mit räumlicher Statistik anschauen:

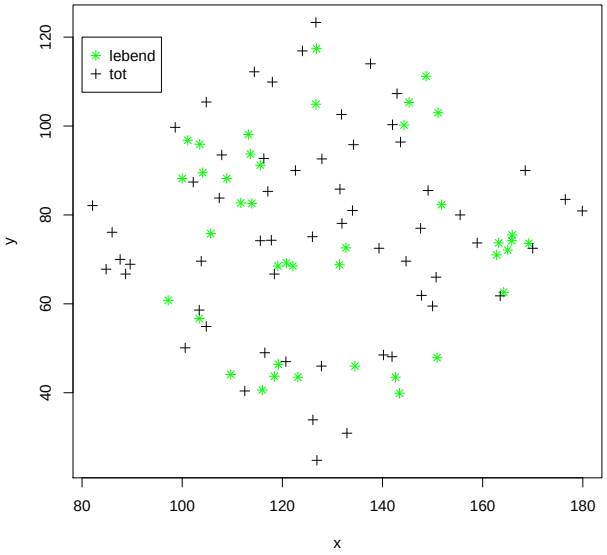


Abbildung 6: 100 tote und lebende Stieleichen (Quercus robur) eines Waldes in Ukiola Natural Park (Baskenland, Nordspanien).

- Es handelt sich um den Datensatz `quercusvm`, der Teil eines größeren Datensatzes, analysiert in Ayastuy [2008]), enthalten im R-package `ecespa` (de la Cruz Rot [2008]), ist.
- Es liegen räumliche Daten zu insgesamt 100 toten und lebenden Stieleichen (Quercus robur) eines Waldes im Urkiola Natural Park[6] (Baskenland, Nordspanien) vor.
- Eine statistisch relevante Fragestellung wäre hier beispielsweise, ob die räumliche Verteilung von toten und lebenden Stieleichen systematisch (und statistisch signifikant) verschieden ist.
- Eine Möglichkeit eines statistischen Tests wäre eine Verallgemeinerung des Kolmogorov-Smirnov-Tests oder des Cramér-von-Mises-Tests für zweidimensionale Verteilungen, siehe beipielsweise Syrjala [1996]. (Natürlich gibt es auch modernere Methoden, etwa könnte man die bivariaten Dichten der Verteilung der lebenden wie der toten Stieleichen beispielsweise mit Kernmethoden schätzen und dann die Differenz der Dichten betrachten.)
- Bei einer Verallgemeinerung des Kolmogorov-Smirnov-Tests gemäß Peacock [1983] würde man zunächst für ein rechteckiges Gebiet B die Differenz D_B zwischen dem Anteil an lebenden Eichen im Bereich B an allen insgesamt 41 lebenden Eichen und dem Anteil von toten Eichen im Bereich B an allen insgesamt 59 toten Eichen berechnen. Anschließend würde man das Supremum
$$D := \sup_{B \text{ rechteckiges Gebiet}} D_B$$
als Teststatistik verwenden.
- Da die Wahl des unterliegenden Koordinatensystems das konkrete Aussehen des Systems aller betrachteten rechteckigen Gebiete beeinflusst, ist die Teststatistik im Allgemeinen abhängig von der Wahl des Koordinatensystems.

[6] `https://en.wikipedia.org/wiki/Urkiola_Natural_Park`

- Ein solches Vorgehen ist also stark von einem analytischen Zugang zur Geometrie geprägt. Während man sich natürlich auch bei einem analytischen Zugang zur reinen Geometrie darum kümmert, dass alle betrachteten Begriffsbildungen nicht von der konkreten Wahl des Koordinatensystems abhängen, ist dieser Anspruch bei obigem Vorgehen zur statistischen Datenanalyse leider nicht mehr erfüllbar.

- Man kann nun fragen, was ein synthetisch geprägtes Verständnis von Geometrie beitragen kann.

- Unser formal-begriffsanalytisches Vorgehen würde zum Betrachten von allen konvexen Gebieten führen, was selbstverständlich kein Koordinatensystem benötigt. (Man könnte meinen, dass wir zur Definition eine Halbraumes an sich bereits auf einen analytischen Zugang angewiesen wären. Dies ist aber nicht so: Wir haben nur auf ein Skalarprodukt zurückgegriffen, das selbstverständlich unabhängig von der Wahl des Koordinatensystems ist. Darüber hinaus hätten wir auch nicht mit Halbräumen, sondern allgemeiner mit konvexen Mengen arbeiten können, und der Begriff der Konvexität lässt sich in einem rein synthetischen Sinne beschreiben: Man kann eine Teilmenge von $\mathbb{R}^2$ einfach als konvex definieren, falls mit je 2 Punkten auch jeder Punkt zwischen diesen 2 Punkten in dieser Menge enthalten ist, was einem rein relationalen Zugang im Sinne von Hilberts Axiomatisierung des Aspekts der Anordnung von Punkten in der Ebene entspricht.

- Die Frage wäre nun, ob das Betrachten aller konvexen Gebiete computational umsetzbar ist und ob die sich ergebende Supremumsstatistik aus rein statistischer Sicht überhaupt irgendetwas leisten kann. Man beachte, dass das Mengensystem aller konvexen Gebiete im Zweifesfall sehr groß ist, und dass die Supremumsstatistik daher ziemlich ins "arg-verteilte" geraten kann, siehe den späteren Abschnitt zur Extremaltheorie von Begriffsverbänden, zur Vapnik-Chervonenkis-Theorie und zur „Regularisierung". Es sei aber schonmal vorweggenommen, dass all diesen Bedenken in einem gewissen Rahmen einigermaßen begegnet werden kann. Die folgende Graphik zeigt eine Auswahl von insgesamt 20 Eichen (in blau), die derart sind, dass keine Eiche in der konvexen Hülle der restlichen der 20 Eichen liegt.

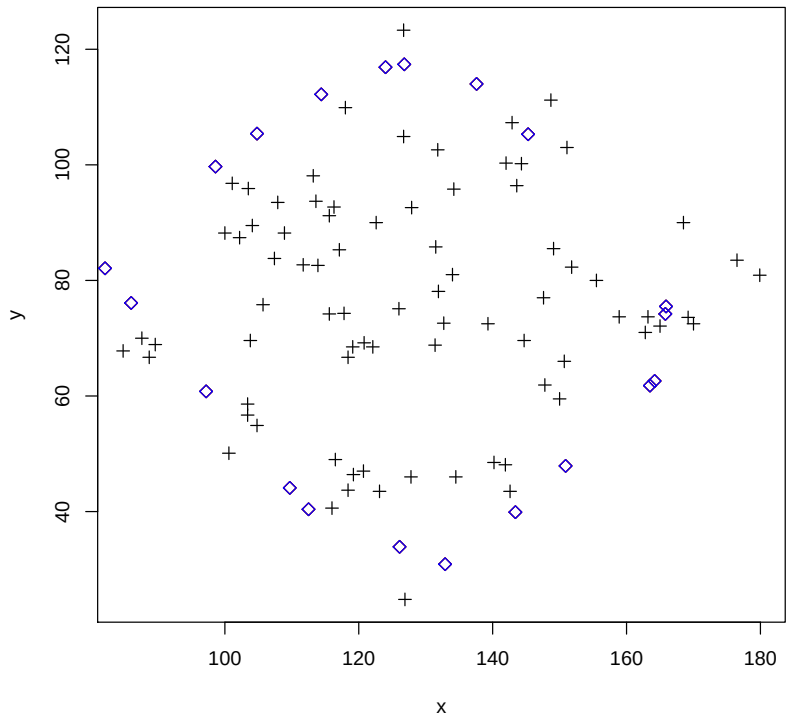
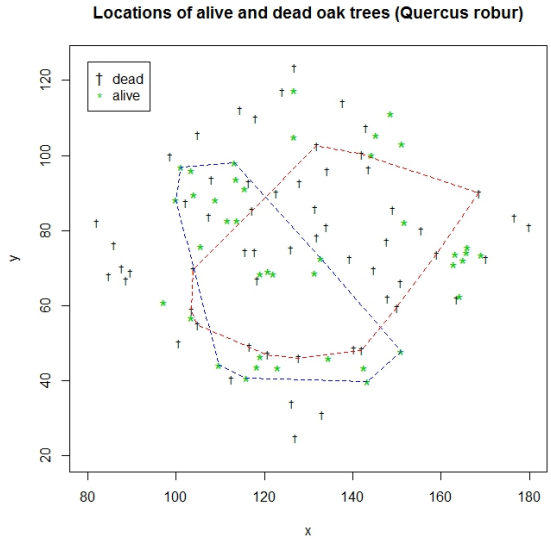


Abbildung 7: In blau: Eine Menge von 20 Stieleichen, die so angeordnet sind, dass keine Eiche in der konvexen Hülle der anderen Eichen liegt.

Es gibt also mindestens $2^{20} = 1.048.576$ formale Begriffe, was noch nicht so viel ist, allerdings ist dies ja nur eine untere Schranke. Eine obere Schranke werden wir noch kennen lernen. Diese würde hier $\binom{100}{20} \approx 5.4 \cdot 10^{20}$ betragen und somit wenig aussagekräftig sein. (Eine konkrete Schätzung der Anzahl aller Begriffe ergibt hier eine Größenordnung von etwa 10^{10}.) Trotzdem kann man hier die Supremumsstatistik berechnen ohne alle konvexen Mengen explizit anschauen bzw. enumerieren zu müssen (vergleiche später). Auch Inferenz über Permutationstests (sowie Regularisierung bzw. auch eine Modifikation des Begriffsverbandes zu einem noch größeren Begriffsverband über eine Verkleinerung der betrachteten Merkmalsimplikationen, siehe später) ist hier noch möglich:



- Implikationsbasis (vgl. später) besitzt 297207 Implikationen
- A) Blau: Konvexe Menge mit größter Differenz der Anteile von lebenden und toten Stieleichen: 61% lebend, 24% tot, Differenz: 37%
- B) Rot: Konvexe Menge mit größter Differenz der Anteile von toten und lebenden Stieleichen: 54% tot, 15% lebend, Differenz: 39%
- p-Wert für Test von A): etwa 0.83
- vgl. Syrjala's Test (zweiseitig):

```
Cramer-von Misses test for the difference between
the spatial distributions of  wx and wy
based on 1e+06 permutations.

   psi:     0.3286233
   p-value: 0.8061132

Kolmogorov-Smirnov test for the difference between
the spatial distributions of  wx and wy
based on 1e+06 permutations.

   psi:     0.168148
   p-value: 0.6656223
```

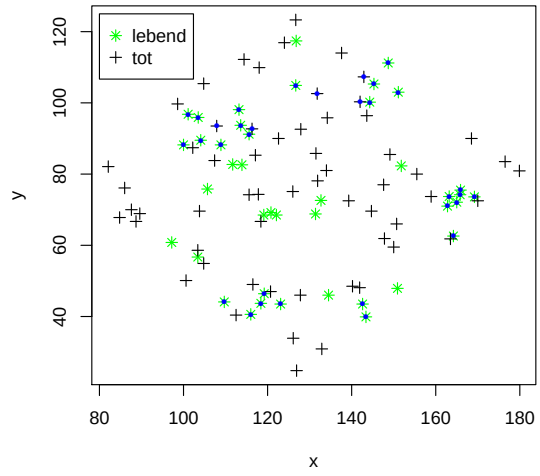


Abbildung 8: Blau: Dasjenige modifizierte (jetzt nicht mehr konvexe) „Gebiet“, für das die Differnez der Anteile von lebenden und toten Stieleichen maximal ist.

Mögliche Modifikation:

- Betrachte nur Implikationen bei denen die beteiligten Prämissen nah genug beieinander liegen, hier konkret Prämissen mit paarweisem Abstand kleinergleich 43 m, dies ergibt hier 132622 Implikationen.
- Blau: „Gebiet“ mit größter Differenz der Anteile von lebenden und toten Stieleichen: 65.9% lebend, 8.5% tot, Differenz: 57.4%
- p-Wert: etwa 0.06

Abschließend seien zum Vergleich noch bivariate Kerndichteschätzungen der Verteilungen von lebenden und toten Stieleichen betrachtet. Eine Schätzung mit einem bivariaten Gaußkern und per Augenmaß gewählter Bandweite ergibt folgende Schätzungen: (Hier wurde die Funktion `bivariate.density` des R-Packages `sparr` mit `h0=5` (ohne adaptive Bandweitenvariation) benutzt.)

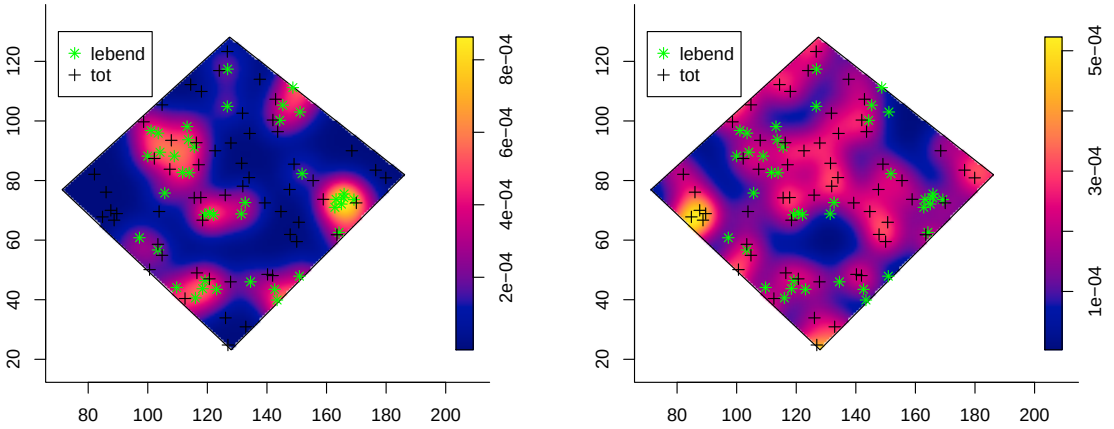


Abbildung 9: Bivariate Kerndichteschätzung der Verteilung der lebenden (links) und toten (rechts) Stieleichen.

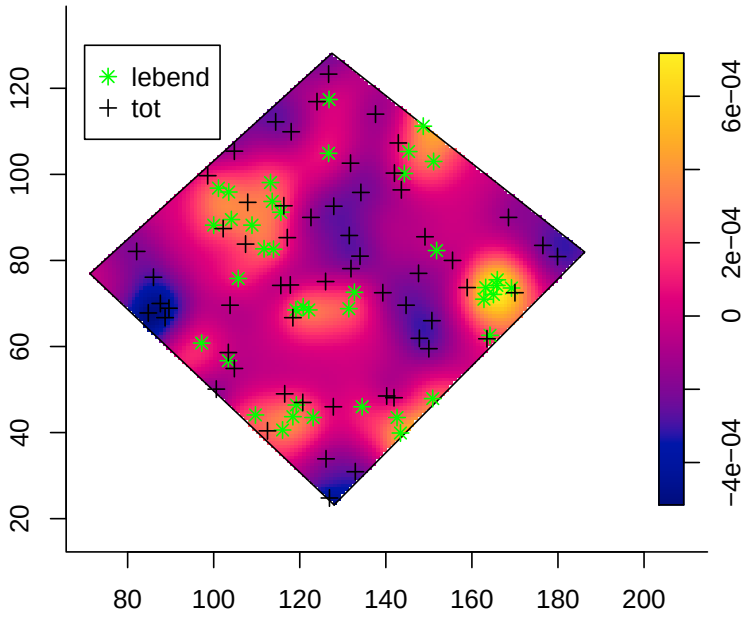


Abbildung 10: Differenz der geschätzten Dichten der Verteilungen der lebenden und der toten Stieleichen.
Ein Permutationstest für die maximale Differenz der Dichten würde hier einen p-Wert von etwa 0.07 ergeben.

Damit wollen wir es mit Beispielen zu Begriffsverbänden auf sich beruhen lassen und kommen zu ein paar allgemeinen **Bemerkungen:**

i) Für einen Gegenstand $g \in G$ bezeichnen wir den von der einelementigen Menge $\{g\}$ erzeugten Begriff

$$(\{g\}'', \{g\}')$$

als **Gegenstandsbegriff**.

ii) Für ein Merkmal $m \in M$ bezeichnen wir den von $\{m\}$ erzeugten Begriff

$$(\{m\}', \{m\}'')$$

als **Merkmalsbegriff**.

iii) Alle Begriffsinhalte entstehen als Schnitte von Begriffsinhalten von Gegenstandsbegriffen: Für einen formalen Begriff (A, B) gilt

$$B = \bigcap_{g \in A} \{g\}'.$$

iv) Alle Begriffsumfänge entstehen als Schnitte von Begriffsumfängen von Merkmalsbegriffen. Für einen formalen Begriff (A, B) gilt

$$A = \bigcap_{m \in B} \{m\}'.$$

v) Alle Begriffsinhalte sind genau die Bilder des mengentheoretischen Hüllenoperators $\Psi \circ \Phi$ auf $(2^M, \subseteq)$.

vi) Alle Begriffsumfänge sind genau die Bilder des mengentheoretischen Kernoperators $\Phi \circ \Psi$ auf $(2^G, \supseteq)$ bzw. des mengentheoretischen Hüllenoperators $\Phi \circ \Psi$ auf $(2^G, \subseteq)$.

vii) Also bildet die Menge aller Begriffsinhalte ein mengentheoretisches Hüllensystem auf $(2^M, \subseteq)$, d.h., das System aller Begriffsinhalte enthält M und ist abgeschlossen unter beliebigen Schnitten.

viii) Beachte: In $(2^M, \subseteq)$ ist das Infimum genau der Mengenschnitt und das Supremum ist genau die Mengenvereinigung. **Warnung:** In $(\mathcal{S}, \subseteq)$ mit $\mathcal{S} \subsetneq 2^M$ ist dies im Allgemeinen nicht so, $(\mathcal{S}, \subseteq)$ muss nicht einmal ein vollständiger Verband sein. (Warum?)

ix) Die Menge aller Begriffsumfänge ist ebenso ein mengentheoretisches Hüllensystem auf $(2^G, \subseteq)$.

x) Jeder beliebige vollständige Verband $(X, \leq)$ kann als Begriffsverband (nämlich genau als $\mathfrak{B}(X, X, \leq)$) dargestellt werden.

xi) Frage: Kann man jedes mengentheoretische Hüllensystem $\mathcal{S} \subseteq 2^M$ auf $(2^M, \subseteq)$ als Hüllensystem von Begriffsinhalten eines geeigneten formalen Kontextes darstellen?

xii) Antwort: Ja. Betrachte dazu den Kontext $\mathbb{K} = (G, M, I)$ mit $G := \mathcal{S}$ und $I := \ni$. Dann gilt für eine beliebige Menge $g \in G = \mathcal{S}$:

$$\{g\}' = \{m \in M \mid gIm\} = \{m \in M \mid g \ni m\} = g,$$

d.h., jedes $g \in \mathcal{S}$ ist ein spezieller Begriffsinhalt. Außerdem ist jeder beliebige Begriffsinhalt als Schnitt von Begriffsinhalten von Gegenstandsbegriffen ebenfalls wieder eine Menge aus $\mathcal{S}$, denn $\mathcal{S}$ ist als Hüllensystem abgeschlossen unter beliebigen Schnitten.

xiii) Die Menge aller Begriffsinhalte ist genau das kleinste Hüllensystem $\mathcal{HS}$ (bezüglich $\subseteq$), das alle Begriffsinhalte von Gegenstandsbegriffen enthält. (Analoge Aussagen gelten jeweils auch für das Hüllensystem aller Begriffsumfänge).

xiv) Frage: Wie kann man dieses Hüllensystem effektiv beschreiben?

xv) Weitere Frage: Wie sieht der kleinste Mengenring $\mathcal{MR}$ aus, der alle Gegenstandsbegriffsinhalte (bzw. Merkmalsbegriffsumfänge) enthält?

xvi) Es gilt $\mathcal{HS} \subseteq \mathcal{MR}$.

2.3 Formale Implikationen

Definition 2.4 (Formale Implikation)
Eine **formale Merkmalsimplikation** $Y \to Z$ ist ein Paar (Y, Z) von Merkmalsmengen und wird interpretiert als Aussage „Alle Gegenstände, die alle Merkmale der Menge Y besitzen, besitzen auch alle Merkmale der Menge Z“. Die Menge Y wird als **Prämisse**, und die Menge Z wird als **Konklusion** bezeichnet. Wir sagen, dass eine formale Implikation $Y \to Z$ in einem formalen Kontext (G, M, I) gilt, falls in der Tat alle Gegenstände des **Kontexts**, die alle Merkmale aus Y besitzen, auch alle Merkmale aus Z besitzen. Eine beliebige Merkmalsmenge B respektiert eine belieibige formale Imlikation $Y \longrightarrow Z$, falls gilt:

$$B \supseteq Y \Longrightarrow B \supseteq Z,$$

d.h, falls B entweder Y nicht vollständig umfasst (also die Prämisse nicht erfüllt) oder falls B mit Y auch Z umfasst.

Eine formale Implikation $Y \to Z$ gilt in einem Kontext genau dann, wenn

$$Y'' \supseteq Z$$

gilt. Außerdem gilt immer

$$Y \longrightarrow Y''.$$

In einem Kontext ist die formale Implikationsrelation $\to$, ebenso, wie die übliche Schlussfolgerungsrelation $\Longrightarrow$ (mit der Interpretation $A \Longrightarrow B$ als: Gelten alle Aussagen der Aussagenmenge A, so folgen logisch notwendig auch alle Aussagen der Menge B) eine **isotone Präordnung**, d.h., es gilt für beliebige $X, Y, Z, W \subseteq M$:

$$\begin{aligned} X \to Y \;\&\; Z \supseteq X &\Longrightarrow Z \to Y && \text{(Isotonie: „Aus mehr folgt mehr“)} \\ X &\to X && \text{(Reflexivität)} \\ X \to Y \;\&\; Y \cup Z \longrightarrow W &\Longrightarrow X \cup Z \longrightarrow W && \text{(Transitivität)} \end{aligned}$$

Bemerkung 2.2. Speziell aus der Transitivität folgt für $Z = \emptyset$, dass

$$X \to Y \;\&\; Y \to W \Longrightarrow X \to W$$

gilt, was der üblichen Definition der Transitivität einer Relation entspricht.

Satz 2.5 (Begriffsinhalte sind genau die respektierenden Mengen)
Die Menge aller Begriffsinhalte eines formalen Kontextes ist genau die Menge aller Menrkmalsmengen, die alle Implikationen, die im Kontext gelten, respektieren.

Beweis.
Sei $\mathcal{A}$ die Menge aller Begriffsinhalte und $\mathcal{B}$ die Menge aller Merkmalsmengen, die alle Implikationen, die im Kontext gelten, respektieren. Wir haben also $\mathcal{A} = \mathcal{B}$ zu zeigen.
Zu $\mathcal{A} \subseteq \mathcal{B}$:

Sei $A \in \mathcal{A}$ und sei $Y \longrightarrow Z$ eine beliebige Implikation, die im Kontext gilt. Wir müssen zeigen, dass A die Implikation $Y \longrightarrow Z$ respektiert, also dass

$$A \supseteq Y \Longrightarrow A \supseteq Z$$

gilt. Sei dazu $B = A'$ der zu A gehörige Begriffsumfang und sei $A \supseteq Y$. Da $A = \bigcap\limits_{g \in B} \{g\}' \supseteq Y$ gilt für jedes $g \in B$, dass $\{g\}' \supseteq Y$ ist. Da die Iplikation $Y \longrightarrow Z$ im Kontext gilt, folgt $\{g\}' \supseteq Z$ für jedes $g \in B$ und somit ist $A \supseteq Z$, d.h., A respektiert die formale Implikation $Y \Longrightarrow Z$.

Zu $\mathcal{B} \subseteq \mathcal{A}$:

Sei $B \in \mathcal{B}$, respektiere also B alle Implikationen, die Im Kontext gelten.
Wir zeigen $B'' = B$, was bedeutet, dass B ein Begriffsinhalt ist. Dazu nehmen wir an, dass $B'' \supsetneq B$ ist und konstruieren daraus einen Widerspruch. Konkret konstruieren wir eine weitere Implikation, die im Kontext gilt, die aber von B nicht respektiert wird:
Es ist

$$B'' = \bigcap \{\{g\}' \mid g \in G : \forall m \in B : gIm\}.$$

Sei nun $b \in B''\backslash B$. Da b aus B'' ist, ist b ein Merkmal, dass alle Gegenstände, die alle Merkmale aus B haben, besitzen. Es gilt also die formale Implikation

$$B \longrightarrow \{b\}.$$

Diese Implikation wird aber offensichtlich nicht von B respektiert, denn es gilt $B \supseteq B$ aber nicht $B \supseteq \{b\}$. Dies ist ein Widerspruch zu $B \in \mathcal{B}$ und somit folgt $B'' = B$, also $B \in \mathcal{A}$.
□

Wenn man sich nun für die formalen Implikationen, die in einem Kontext gelten, interessiert, ist es oft hilfreich, nicht die Menge aller formalen Implikationen, sondern nur die Menge aller „nichttrivialen“ Implikationen (, d.h. ungefähr: derjenigen Implikationen, die nicht schon aus Reflexivität, Isotonie und Transitivität folgen) zu betrachten, da die Menge der formalen Implikationen in konkreten Anwendungssituationen schon sehr groß sein kann.

Definition 2.6 (Implikationsbasis)
Sei $\mathbb{K} := (G, M, I)$ ein formaler Kontext und $\mathcal{I}(\mathbb{K})$ die Menge aller formalen Implikationen, die in $\mathbb{K}$ gelten. Eine Teilmenge $J \subseteq \mathcal{I}(\mathbb{K})$ heißt **Implikationsbasis** für $\mathbb{K}$, falls gilt:

$$H(J) := \bigcap \{R \mid R \subseteq 2^M \times 2^M, R \supseteq J, R \text{ ist isotone Präordnung auf } 2^M\} = \mathcal{I}(\mathbb{K})$$

und falls zusätzlich J bezüglich dieser Eigenschaft minimal ist, d.h., falls für jede echte Teilmenge $L \subsetneq J$ bereits $H(L) \subsetneq \mathcal{I}(\mathbb{K})$ gilt.

Bemerkung 2.3. Wegen $Y \to Z \iff Y'' \supseteq Z$ ist alles, was formal aus Y folgt, genau gleich Y''. Man erhält also für jedes $Y \subseteq M$ die formale Implikation $Y \to Y''$, die man noch um ihren trivialen Teil $Y \longrightarrow Y$ zu $Y \to Y'' \backslash Y$ kürzen kann. Falls $Y'' = Y$ gilt, kann man die zugehörige triviale Implikation $Y \to Y \backslash Y$ einfach weglassen.

Bemerkung 2.4. Dual zu Merkmalsimplikationen kann man sich auch Gegenstandsimplikationen anschauen, indem man einfach die Rolle von Merkmalen und Gegenständen vertauscht.

Bemerkung 2.5. Im Allgemeinen gibt es natürlich mehrere Implikationsbasen.

Bemerkung 2.6. Die Berechnung einer Implikationsbasis, konkret der sogenannten Stammbasis kann auch hier wieder über die Berechnung eines geeigneten Hüllensystems, konkret des Hüllensystems aller Begriffsinhalte und aller sogenannten Pseudoinhalte (siehe [Ganter, 1996, Abschnitt 2.3, insbesondere ab S. 83]) geschehen. Dieses Hüllensystem kann allerdings im Allgemeinen wieder sehr groß sein. Insbesondere ist dieses Vorgehen im Fall, dass man eine kleine Implikationsbasis, aber ein großes System an Begriffsinhalten hat, nicht sehr effektiv. (Vergleiche auch nachfolgendes Beispiel zu konvexen Mengen in $\mathbb{R}^2$.)

Bemerkung 2.7. Bei obigem Algorithmus werden nach und nach Implikationen generiert. Das konkrete Vorgehen des Algorithmus macht es möglich, eine sogenannte Merkmalexploration durchzuführen. Dabei wird der Benutzer für jede generierte Implikation der Implikationsbasis gefragt, ob diese nur kontingenterweise im Kontext gilt, oder ob diese auch schlechthin gilt. Bejat der Nutzer diese Frage, dann wird die Implikation in die Basis aufgenommen und es wird die nächste Implikation generiert. Verneint der Benutzer die Frage, dann muss er ein Gegenbeispiel angeben, das zeigt, dass die Implikation nicht gilt. Dieses Beispiel wird dann mit in den Kontext aufgenommen. Die Abfolge der Generation der Implikationen stellt sicher, dass zuvor generierte und als schlechthin geltend bewertete Implikationen im späteren Verlauf immer noch gültig bleiben und nicht erneut berechnet werden müssen. Außerdem werden dem Benutzer nie Implikationen präsentiert, die aus den vorher als geltend bewerteten Implikationen bereits logisch folgen, d.h., die interaktive Abfrage der Gültgkeit von Implikationen ist sehr sparsam gehalten.

Kommen wir nun zu einem Beispiel, wo eine Implikationsbasis relativ einfach explizit zu beschreiben ist:

Beispiel 17. Betrachte wieder
$G = \mathbb{R}^2 \ldots$ Menge aller Punkte in $\mathbb{R}^2$,

$M = \{\{x \in \mathbb{R}^2 \mid \langle x, v\rangle < c\} \mid v \in \mathbb{R}^2, c \in \mathbb{R}\} \ldots$ Menge aller offenen Halbräume von $\mathbb{R}^2$,

$gIm \iff$ Punkt g liegt in Halbraum m.

Die Begriffsumfänge waren dann die konvexen Mengen von $\mathbb{R}^2$. Wie sehen dann die Gegenstandsimplikationen aus?

i) Implikationen mit der leeren Menge als Prämisse („triviale“ Implikationen): Es gelten keinerlei triviale Implikationen.

ii) Einfache Implikationen, also Implikationen mit einelementiger Prämisse: Da gelten nur die redundanten Implikationen $\{x\} \longrightarrow \{x\}$.

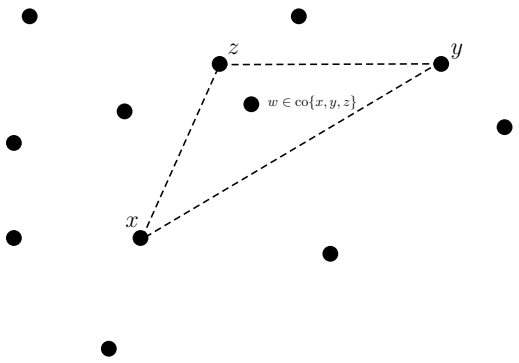
iii) Implikationen mit zweielementiger Prämisse: Für unterschiedliche $x, y \in \mathbb{R}^2$ gilt: $\{x, y\} \longrightarrow \{\lambda \cdot x + (1-\lambda) \cdot y \mid \lambda \in [0,1]\}$.

iv) Implikationen mit dreielementiger Prämisse: Für paarweise verschiedene $x, y, z \in \mathbb{R}^2$ gilt: $\{x, y, z\} \longrightarrow \{\lambda \cdot x + \mu \cdot y + \nu \cdot z \mid \lambda, \mu, \nu \in [0,1], \lambda + \mu + \nu = 1\}$.

v) Implikationen mit k-elementiger Prämisse: Für paarweise verschiedene $x^1, x^2, \ldots, x^k \in \mathbb{R}^d$ gilt: $\{x^1, x^2, \ldots, x^k\} \longrightarrow \left\{ \sum_{i=1}^{k} \lambda_i \cdot x^i \mid \lambda_1, \lambda_2, \ldots, \lambda_k \in [0,1], \sum_{i=1}^{k} \lambda_i = 1 \right\}$.

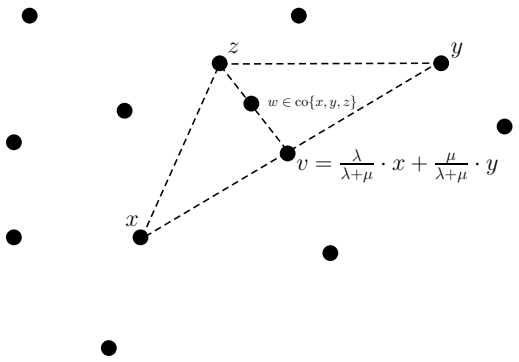
vi) Allgemein für beliebige Prämissen Y gilt $Y \longrightarrow \mathrm{co}\,(Y)$, wobei $\mathrm{co}\,(Y)$ die konvexe Hülle von Y ist.

Und wie könnte eine Implikationsbasis aussehen? Beobachtung: Für eine beliebige Menge Y von Punkten gilt: Jeder Punkt aus der konvexen Hülle von Y liegt bereits in der konvexen Hülle von 3 Punkten aus Y. Dies ist genau der Satz über konvexe Mengen von Carathéodory für $d = 2$.



Damit reicht es aus, Implikationen mit zwei- bzw. dreielementigen Prämissen zu betrachten. Für $G = \mathbb{R}^2$ kann man sich leicht überlegen, dass Implikationen mit dreielementiger Prämisse (und auch Implikationen mit mehr als 3 Elementen in der Prämisse) redundant sind, da sie bereits durch geeignet gewählte Implikationen mit zweielementiger Prämisse abgedeckt werden: Für $x, y, z \in \mathbb{R}^2$ und $\lambda, \mu, \nu \in [0,1]$ mit $\lambda+\mu+\nu = 1$ kann die Konvexkombination $\lambda \cdot x + \mu \cdot y + \nu \cdot z$ dargestellt werden als

$$\lambda \cdot x + \mu \cdot y + \nu \cdot z = (\lambda + \mu) \cdot \left(\frac{\lambda}{\lambda + \mu} \cdot x + \frac{\mu}{\lambda + \mu} \cdot y \right) + \nu \cdot z.$$

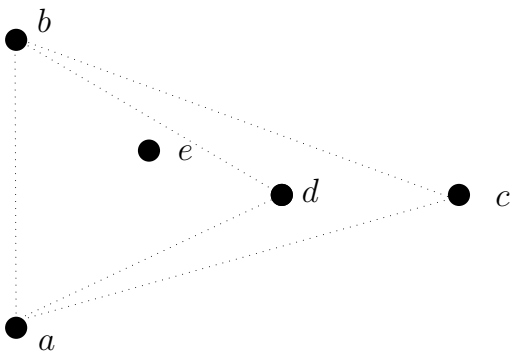


Da für eine konvexe Menge A mit den Punkten x und y auch der Punkt $v = \left(\frac{\lambda}{\lambda+\mu} \cdot x + \frac{\mu}{\lambda+\mu} \cdot y\right)$ vermöge der zugehörigen Implikation $\{x, y\} \longrightarrow \mathrm{co}\,(\{x, y\})$ wieder in A liegen muss (beachte $\frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} = 1$), und da $(\lambda + \mu) + \nu = 1$ ist, muss schließlich auch $\lambda \cdot x + \mu \cdot y + \nu \cdot z$ bereits aufgrund von Implikationen mit zweielementiger Prämisse ebenso wieder in A liegen.

Für den Fall, dass endlich viele Punkte in $\mathbb{R}^2$ als Gegenstandsmenge betrachtet werden, ist die Situation ein wenig komplizierter. Befinden sich die Punkte in allgemeiner Lage, d.h., liegt kein Punkt genau zwischen zwei anderen Punkten, so ist eine Implikationsbasis gegeben durch

$$J := \{Y \longrightarrow \mathrm{co}\,(Y) \cap G \mid Y \subseteq \mathbb{R}^2, |Y| = 3\}.$$

Allgemein ist hier ein bisschen Vorsicht geboten. Beipielsweise für $G = \{a, b, c, d, e\}$ gemäß unten skizzierte Situation gelten unter Anderem die folgenden Implikationen mit dreielementigen Prämissen:



$$\{a, b, c\} \longrightarrow \{d, e\}$$
$$\{a, b, d\} \longrightarrow \{e\}$$

und diese sind auch nicht redundant. Betrachtet man anstelle obiger Implikationen jedoch die Implikationen

$$\{a, b, c\} \longrightarrow \{d\}$$
$$\{a, b, c\} \longrightarrow \{e\}$$
$$\{a, b, d\} \longrightarrow \{e\},$$

die entstehen, indem man die Konklusionen aufspaltet, dann ist hier konkret die Implikation $\{a, b, c\} \longrightarrow \{e\}$ redundant, da sie für jede isotone Präordnung aus den anderen beiden Implikationen über die Transitivität folgt.

Abschließend sei das folgende **Ergebnis** festgehalten:

Für eine endliche Anzahl von n Punkten in $\mathbb{R}^2$ (in Verbindung mit der Inzidenz zu Halbräumen in $\mathbb{R}^2$) benötigt man für die Beschreibung des zugehörigen Begriffsverbands schlimmstenfalls 2^n Begriffsumfänge. Demgegenüber gibt es eine Beschreibung des Begriffsverbands mit einer Implikationsbasis, die eine Größenordnung von n^3 besitzt!

2.4 Begriffliches Skalieren

Bisher hatten wir immer nur dichotome Merkmale betrachtet, d.h., ein Gegenstand konnte immer nur ein Merkmal besitzen oder nicht. Im Zusammenhang mit statistischer Datenanalyse hat man aber oft auch mit Merkmalen zu tun, die ein höheres Skalenniveau aufweisen, die beispielsweise nominal oder ordinal skaliert sind. Es stellt sich daher die Frage, ob die formale Begriffsanalyse auch in solchen Situationen angewendet werden kann? Die Antwort ist (zumindest in einem konzeptionellen Sinne) ein klares „Ja und weit darüber hinaus!“: Beinahe jedem Skalenniveau kann hier jedenfalls im Prinzip Rechnung getragen werden. Dies geht weit über die Stevenssche Skalenhirarchie hinaus! Die Grundidee beim sogenannten begrifflichen Skalieren ist einfach die, ein komplexes Merkmal in eine Menge von dichotomen Merkmalen zu zerlegen. In bestimmten Situationen gibt es dazu meist mehrere denkbare und auch sinnvolle Möglichkeiten. Dies ist genau eine der Stärken der formalen Begriffsanalyse, man kann immer versuchen, dem Charakter des betrachteten Merkmals „gerecht zu werden“, man muss nicht die zur Verfügung stehende Methode das Skalenniveau diktieren lassen. Dabei ist es natürlich wichtig, sich immer genau zu überlegen, wie der bei bestimmter begrifflicher Skalierung entstehende Begriffsverband aussieht und ob die entstehenden formalen Begriffe angemessen sind, um die entsprechend anvisierte Fragestellung zu betrachten. Dies erfordert sicherlich meist eine Art Werturteil, aber dies ist eben letztendlich genau eine Stärke der formalen Begriffsanalyse, sie fordert uns auf, unsere

„unbewußten Zwecke auf[zu]decken, [unsere] ... bewußten Zwecke [zu] deklarieren,“ und unsere *„Mittel danach aus[zu]wählen und aus[zu]richten...,“* [von Hentig 1974, S. 136]

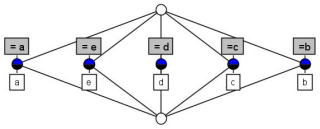
damit wir *„...wissen, was wir denken, ...“* [Peirce, 1878]

„ ...damit wir Herren unserer eigenen Meinung werden, Gedanken ohne Bedeutung vermeiden und unsere Sprache eine klare werde.“[Oehler, S. 102 in der Besprechung von Peirce [1878]]

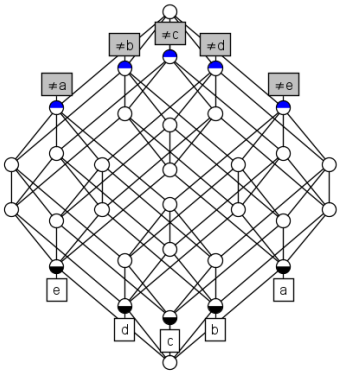
Wir betrachten nun ein paar **Beispiele:**

- Nominalskalierte Merkmale (z.B. $x \in \{a, b, c, d, e\}$) mit Nominalskala:

	=a	=b	=c	=d	=e
a	×				
b		×			
c			×		
d				×	
e					×



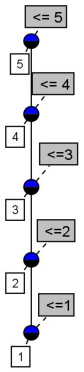
- Nominalskalierte Merkmale (z.B. $x \in \{a, b, c, d, e\}$) mit Kontranominalskala:



	≠a	≠b	≠c	≠d	≠e
a		×	×	×	×
b	×		×	×	×
c	×	×		×	×
d	×	×	×		×
e	×	×	×	×	

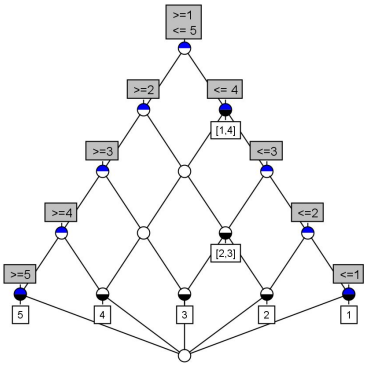
- Ordinalskalierte Merkmale (z.B. $x \in \{1, 2, 3, 4, 5\}$) mit Ordinalskala:

A	B	C	D	E	F
	<=1	<=2	<=3	<= 4	<= 5
1	X	X	X	X	X
2		X	X	X	X
3			X	X	X
4				X	X
5					X



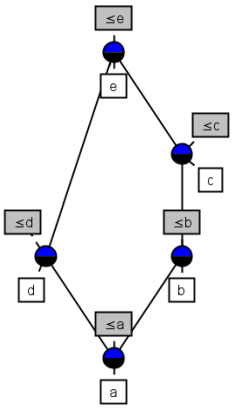
- Ordinalskalierte Merkmale (z.B. $x \in \{1, 2, 3, 4, 5\}$) mit Interordinalskala. Hier können auch intervallwertige Beobachtungen behandelt werden:

A	B	C	D	E	F	G	H	I	J	K
	<=1	<=2	<=3	<= 4	<= 5	>=1	>=2	>=3	>=4	>=5
1	X	X	X	X	X	X				
2			X	X	X	X	X			
3				X	X	X	X	X		
4					X	X	X	X	X	
5						X	X	X	X	X
[2,3]			X	X	X	X	X			
[1,4]										



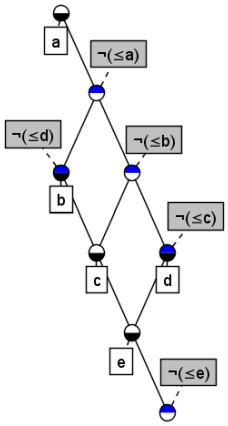
- Partiell geordnete Merkmale (z.B. $x \in \{a, b, c, d, e\}$ mit $\lessdot := \{(a, b), (b, c), (c, e), (a, d), (d, e)\}$), skaliert mit Ordinalskala:

	≤a	≤b	≤c	≤d	≤e
a	X	X	X	X	X
b		X	X		X
c			X		X
d				X	X
e					X



- Partiell gerodnete Merkmale (z.B. $x \in \{a, b, c, d, e\}$ wieder mit $\lessdot := \{(a, b), (b, c), (c, e), (a, d), (d, e)\}$). Jetzt aber mit Kontraordinalskala, d.h. $I := \not\leq$:

	¬(≤a)	¬(≤b)	¬(≤c)	¬(≤d)	¬(≤e)
a					
b	×			×	
c	×	×		×	
d	×	×	×		
e	×	×	×	×	



Bemerkung: Die Begriffsumfänge sind genau die Oberhalbmengen und die Begriffsinhalte sind genau die Unterhalbmengen von $(X, \leq)$.

- Ein weiteres konkretes Beispiel: Analyse von Ranking Daten im Kontext des **wisdom of the crowd** Phänomens, siehe Lee et al. [2014]: 26 Personen wurden gebeten, die damals noch 44 US-Präsidenten entsprechend der vermuteten zeitlichen Reihenfolge ihrer Präsidentschaft zu ranken. Also haben wir 26 Datenpunkte, wobei jeder Datenpunkt eine totale Ordnung auf einer 44-elementigen Grundmenge ist, man hat also ordnungswertige Daten. Auch solche Daten kann man einfach begrifflich skalieren, indem man für jedes Paar (p, q) von zwei verschiedenen Präsidenten das binäre Merkmal „die befragte Person rankt Präsident p zeitlich nach Präsident q“ einführt:

A	B	C	D	E	F
	Thomas Jefferson zeitlich nach John Adams	James Madison zeitlich nach John Adams	James Monroe zeitlich nach John Adams	John Quincy Adams zeitlich nach John Adams	Andrew Jackson zeitlich nach John Adams
1					×
2			×		
3	×	×		×	×
4	×		×	×	×
5	×		×	×	×
6	×	×	×	×	×
7	×	×	×	×	×
8	×	×	×	×	×
9		×	×	×	×
10		×	×	×	×
11	×	×	×	×	×
12	×	×	×	×	×
13	×	×	×	×	×
14	×	×	×	×	×
15	×	×	×	×	×
16	×	×	×	×	×
17	×	×	×	×	×
18	×	×	×	×	×
19	×	×	×	×	×
20					×
21	×	×	×	×	×
22	×	×	×	×	×
23	×	×	×	×	×
24	×	×	×	×	×
25		×	×	×	×
26	×	×	×	×	×

...

ASZ	ATA	ATB	ATC
William McKinley zeitlich nach James Carter	Theodore Roosevelt zeitlich nach James Carter	William Howard Taft zeitlich nach James Carter	Woodrow Wilson zeitlich nach James Carter
	×		
	×		×
×		×	×
×	×	×	
	×		
×	×	×	×
	×	×	×
			×
×		×	
			×

...

Insgesamt hätten wir also $44 \cdot 43 = 1892$ Merkmale. Da wir aber nur 26 Personen, also 26 Gegenstände haben, gibt es maximal 2^{26} formale Begriffe, konkret hat der zugehörige Begriffsverband insgesamt 10.452.368 formale Begriffe, so dass eine graphische Darstellung hier nicht mehr sinnvoll ist. Im nächsten Abschnitt sollen zunächst zwei ausgewählte Möglichkeiten betrachtet werden, um mit großen Begriffsverbänden umzugehen. (Es gibt darüber hinaus noch viele weitere Möglichkeiten, ein paar weitere Möglichkeiten, die von einem statistischen bzw. kombinatorischen Zugang geprägt sind, folgen später noch. Insbesondere werden wir auch noch im Abschnitt zur Extremaltheorie für Begriffsverbände explizit anschauen, was Begriffsverbände besonders groß macht.)

2.5 Was, wenn $\mathfrak{B}((G, M, I))$ sehr groß ist?

2.5.1 Iceberg Begriffsverbände

Abbildung 11: Eisberg (Fotomontage) Created by Uwe Kils (iceberg) and User:Wiska Bodo (sky). - (Work by Uwe Kils) http://www.ecoscope.com/iceberg/, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=209674

- Nur sehr "interessante" Begriffe anschauen, dies sind in einigen Situationen vor allem Begriffe mit einem sehr großen Begriffsumfang.
- Betrachte nur Begriffe, deren Begriffsumfänge eine Mächtigkeit ("**support**") größergleich einer gewissen Mindestgröße (**minsupp**) haben. ("Betrachte nur die Spitze des riesigen Eisbergs.")
- Wie können nur die formalen Begriffe mit **support** größergleich einer gegebenen Mindestschranke **minsupp** berechnet werden, ohne vorher alle Begriffe berechnen zu müssen?
- Beobachtung: Für Merkmalsmengen B_1, B_2 mit $B_1 \subseteq B_2$ folgt $B_1' \supseteq B_2'$, dies bedeutet insbesondere: Aus $|B_1'| <$ **minsupp** folgt $|B_2'| <$ **minsupp**. Also: Gilt für eine Merkmalsmenge B die Relation $|B'| <$ **minsupp**, so gehört B nicht mehr zur "Spitze des Eisbergs" und darüber hinaus müssen keine weiteren Obermengen von B betrachtet werden, da diese erst recht nicht mehr zur Spitze gehören.

2.5.2 Quantile

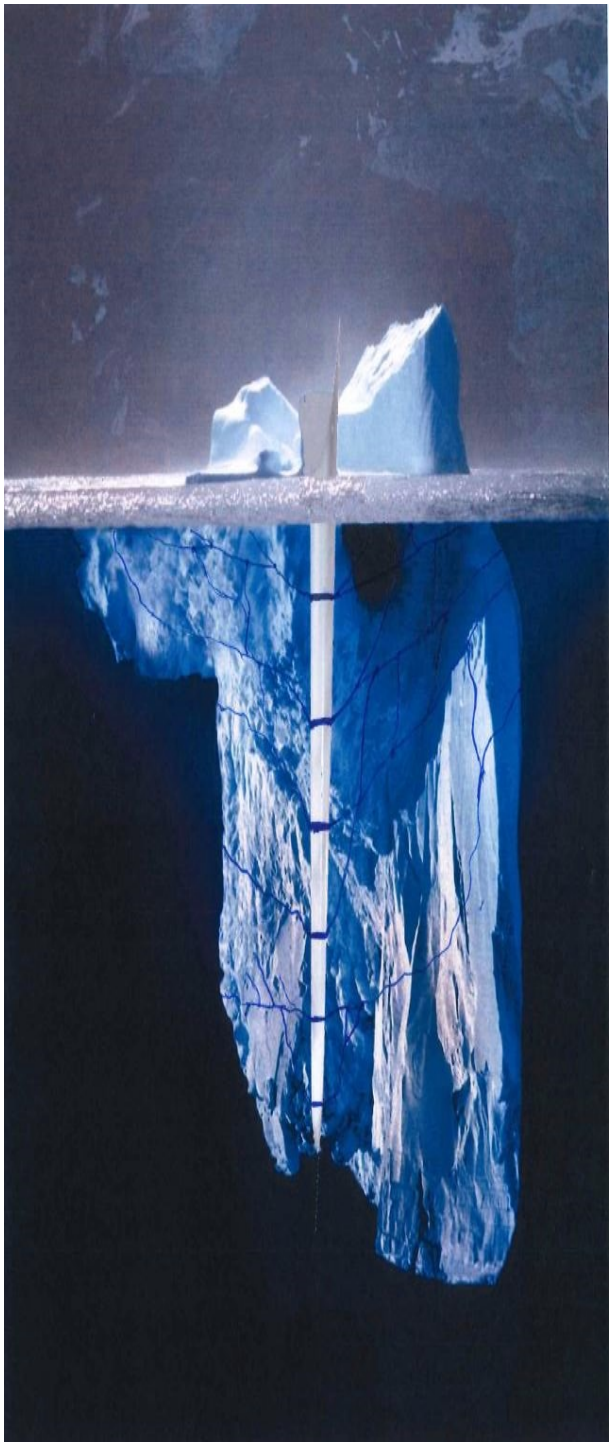


Abbildung 12: Eisberg mit eingezeichnetem Querschnitt (Bearbeitung von Abbildung 11)

- Betrachte keinen "horizontalen Schnitt durch den Begriffsverband" wie bei den Iceberg Begriffsverbänden geschehen, sondern betrachte einen vertikalen Schnitt von "repräsentativen" Begriffen mit aufsteigenden Begriffsumfängen (bzw. absteigenden Begriffsinhalten).
- Problem: Wie soll man für gegebenen **support** einen repräsentativen Begriff mit entsprechendem support auswählen?
- Antwort: indirekt:
  1. Betrachte zunächst Anteil $\alpha \in [0, 1]$.
  2. Betrachte dann die Menge M_α **aller** formalen Begriffe zu denen mindestens $\alpha \cdot |G|$ Gegenstände gehören.
  3. Wähle dann den größten gemeinsamen Unterbegriff $x_\alpha = \bigwedge M_\alpha$ aller dieser Begriffe als einen repräsentativen Begriff.
  4. Zum Begriff $x_\alpha = \bigwedge M_\alpha$ gehören dann im Allgemeinen weniger als $\alpha \cdot |G|$ Gegenstände, was der Konstruktion an sich aber nichts zur Sache tut.
  5. Beobachtung: Aus $\alpha \leq \beta$ folgt $M_\alpha \supseteq M_\beta$ und somit $x_\alpha = \bigwedge M_\alpha \leq \bigwedge M_\beta = x_\beta$, so dass für verschiedene Anteile α, β die zugehörigen Begriffe immer in Relation stehen. Konkret ist für $\alpha \leq \beta$ der Begriff x_α ein Unterbegriff von x_β. Die Menge aller Quantile ist also eine total geordnete Teilmenge aller Begriffe.
  6. Zu jedem Gegenstand g (bzw. allgemeiner zu jeder Gegenstandsmenge A) kann man dann in natürlicher Weise den spezifischsten Begriff, der diesen Gegenstand g (bzw. die Gegenstände der betrachteten Menge A) enthält, betrachten.

Bemerkung 2.8. Da jeder vollständige Verband als Begriffsverband vorstellbar ist, kann obige Idee auch auf Datensätze (bzw. auch Zufallsvariablen) mit Werten in einem vollständigen Verband übertragen werden. Das einzige was nicht ganz offensichtlich ist, ist die Antwort auf die Frage, wie man genau M_α definieren soll, da hier ja konkret der Anteil $\frac{|A|}{|G|}$ eingeht, und dieser ist bei Datensätzen/Zufallsvariablen mit Werten in einem beliebigen vollständigen Verband nicht a priori gegeben. Das natürlichste wäre wohl, M_α als Menge aller x, unterhalb derer

mindestens $\alpha \cdot |G|$ Datenpunkte liegen, zu definieren. Für eine Zufallsvariable Z mit Werten in einem vollständigen Verband $(X, \leq)$ kann man analog M_α definieren als Menge aller $x \in X$ mit $P(Z \leq x) \geq \alpha$.

Frage: Muss man zur Berechnung von x_α alle Begriffe ausrechnen? Antwort: Nein, es reicht, lediglich alle $|M|$ Merkmalsbegriffe anzuschauen: Sei U_α die Menge aller Begriffsumfänge der Begriffe aus M_α. Sei weiter QU_α der Begriffsumfang des Quantils x_α. Dann ist

$$QU_\alpha = \bigcap U_\alpha$$

gemäß des Hauptsatzes der formalen Begriffsanalyse. Weiter gilt nun

$$\bigcap U_\alpha = \bigcap \left\{ \{b\}' \mid \frac{|\{b\}'|}{|G|} \geq \alpha \right\},$$

d.h., der Umfang des Quantils x_α entsteht bereits als Schnitt aller Umfänge von Merkmalsbegriffen, die eine Mächigkeit von mindestens $\alpha \cdot |G|$ besitzen. Begründung:

„$\subseteq$“: Klar, denn $U_\alpha \supseteq \left\{ \{b\}' \mid \frac{|\{b\}'|}{|G|} \geq \alpha \right\}$.

„$\supseteq$“: Dies zeigen wir indirekt, indem wir $g \notin \bigcap U_\alpha \Longrightarrow g \notin \bigcap \left\{ \{b\}' \mid \frac{|\{b\}'|}{|G|} \geq \alpha \right\}$ zeigen. Sei dazu also $g \notin \bigcap U_\alpha$. Dann existiert ein Begriff (A, B) mit $\frac{|A|}{|G|} \geq \alpha$ und $g \notin A$. Insbesondere muss dann für diesen Begriff ein $b \in B$ existieren mit $g \not I b$. Für dieses b gilt $g \notin \{b\}'$ und $\frac{|\{b\}'|}{|G|} \geq \frac{|A|}{|G|} \geq \alpha$. Also folgt $g \notin \bigcap \left\{ \{b\}' \mid \frac{|\{b\}'|}{|G|} \geq \alpha \right\}$.

Frage: Was bedeutet es, dass ein Gegenstand zum Umfang eines Quantils $x_\alpha = \bigwedge M_\alpha$ gehört bzw. was bedeutet es, dass ein Begriff Unterbegriff eines Quantils x_α ist? Für einen beliebigen Begriff (A, B) gilt: $(A, B) \leq \bigwedge M_\alpha \iff (A, B) \leq (C, D)$ für jeden Begriff $(C, D) \in M_\alpha \iff$ Jeder allgemein genuge Begriff (im Sinne von Begriffsumfang hat Mächtigkeit größergleich $\alpha \cdot |G|$) ist Oberbegriff von (A, B). Bzw. für ein $g \in G$: $(\{g\}'', \{g\}') \leq M_\alpha \iff$ Jeder allgemein genuge Begriff enthält g im Umfang. Betrachte jetzt zu $g \in G$ (bzw. allgemein für $(A, B) \in \mathfrak{B}((G, M, I))$) das spezifischste Quantil

$$\bigwedge M_{\alpha(g)}$$

(im Sinne von α ist minimal), das diesen Gegenstand noch enthält (bzw. gerade noch Oberbegriff von (A, B) ist). Beachte, dass für verschiedene α die zugehörigen Quantile identisch sein können. Außerdem nehmen wir hier an, dass der Begriffsverband endlich ist, so dass es immer ein minimales Quantil bzw. ein minimales α gibt. Was bedeutet es dann in etwa, wenn beispielsweise $\alpha(g)$ sehr groß ist? In etwa vielleicht: Nur sehr allgemeine Begriffe enthalten sicher den Gegenstand g. Beachte aber: Für ein $g \in G$ enthält der Gegenstandsbegriff $(\{g\}'', \{g\}')$ den Gegenstand g. Dieser Begriff ist der spezifischste Begriff, der g enthält und dieser Begriff ist typischerweise sehr spezifisch, aber nicht unbedingt „repräsentativ“ für die gesamte „Population von Gegenständen“. Demgegenüber könnte man die Quantile als „repräsentativ“ bezeichnen in dem Sinne, dass diese nicht mit Referenz auf das Ansehen ausgewählter Gegenstände, sondern durch das schlichte Zählen von nicht näher besehenen Gegenständen und eine anschließende neutrale Begriffskonstruktion (nämlich das Bilden des allgemeinsten Unterbegriffs von M_α) konstruiert wurden. Ist dagegen $\alpha(g)$ sehr klein, so ist g gewöhnlich in dem Sinne, dass viele Quantile g enthalten. Diejenigen Gegenstände, die zum kleinstmöglichen α gehören, könnte man somit auch als die zentralsten Gegenstände bezeichnen. Soviel zur allgemeinen Interpretation der Quantile. Gegeben eine konkrete Datensituation kann (und sollte) die Interpretation

natürlich weiter konkretisiert werden, insbesondere spielt natürlich auch die konkrete begrifflich Skalierung eine Rolle. Es folgen ein paar Beispiele ohne viel Worte:

Beispiel 18. Für ein ordinal skaliertes Merkmal mit ordinaler begrifflicher Skalierung entsteht im Wesentlichen eine asymmetrische Variante des vertrauten (unteren) Quantilkonzepts.

Beispiel 19. Für ein ordinal skaliertes Merkmal mit interordinaler begrifflicher Skalierung entsteht ein klassisches ordinales Outlyingnesskonzept: Beispielweise für 5 beobachtete Datenpunkte $1 < 2 < 3 < 4 < 5$ erhalten wir so etwas:

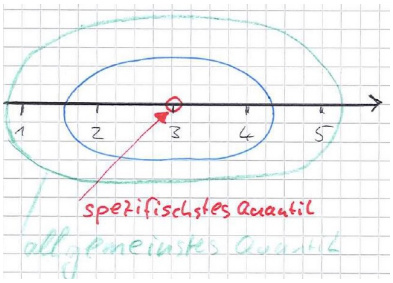


Abbildung 13: Formal-begriffliches Quantilkonzept für ordinale Daten und interordinale begriffliche Skalierung: Ein klassisches ordinales Outlyingnesskonzept.

Beispiel 20. Für (G, M, I) mit $G = \mathbb{R}^2 \cap \{x_1, \ldots, x_n\}$, $M \ldots$ Menge aller Halbräume in $\mathbb{R}^2$ und gIm falls Punkt g in Halbraum m liegt, erhalten wir im Wesentlichen ein ebenso bekanntes Outlyingnesskonzept für $\mathbb{R}^2$, nämlich Tukey's half space depth:

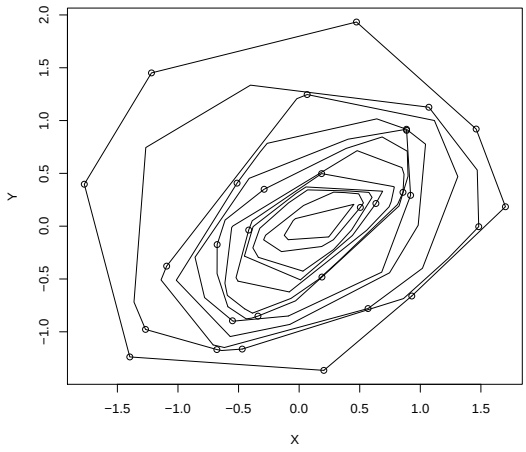


Abbildung 14: Illustration von 30 Datenpunkten in $\mathbb{R}^2$ mit zugehörigen "Tiefenschichten" bei Betrachtung von Tukey's half space depth.

Bemerkung 2.9. In allen bisherigen Beispielen waren die Gegenstände bzw. die Gegenstandsbegriffe als statistische Einheiten gedacht. Es ist natürlich auch möglich, dass die statistischen Einheiten beliebige Begriffe sind. Beispielsweise beim letzten Beispiel könnte man sich leicht vorstellen, dass man keine Punkte in $\mathbb{R}^2$, sondern allgemein konvexe[7] Punktmengen von $\mathbb{R}^2$ beobachtet, man denke beispielsweise an räumlich ausgedehnte Objekte, z.B. Gewässer in einer

[7]Für nicht-konvexe Mengen kann man einfach deren konvexe Hülle betrachten. Im allgemeinen Fall kann man analog für beliebige Gegenstandsmengen den erzeugten Begriffsumfang betrachten.

Landschaft[8]. Dann kann man für $\alpha \in [0,1]$ in natürlicherweise die Menge

$$M_\alpha := \{(A,B) \in \mathfrak{B}((G,M,I)) \,|(A,B) \text{ ist Oberbegriff von mindestens } \alpha \cdot N \text{ von insgesamt } N \text{ beobachteten Begriffen}\}$$

und das zugehörige Quantil $x_\alpha = \bigwedge M_\alpha$ betrachten.

Bemerkung 2.10. Allgemein können wir das Quantilkonzept als ein rein formalbegriffsanalytisches Outlyingnesskonzept auffassen. Wichtig ist hier die Allgemeinheit der Konstruktion: Im Prinzip können wir alles, was wir sinnvoll begrifflich skalieren können, hier behandeln!

Es folgt jetzt noch eine Anwendung des Quantilkonzepts für die Situation obigen Beispiels, die aber auch ganz allgemein betrachtet werden kann:

Anwendung: Kombinatorische „Regularisierung“ eines Kolmogorov-Smirnov-artigen Tests für $\mathbb{R}^2$ (Vergleiche auch den späteren Abschnitt zur Regularisierung)

Gegeben sei eine Population bestehend aus zwei Subpopulationen 1 und 2 (man denke beispielsweise an obiges Beispiel von toten und lebenden Stieleichen) und die Frage, ob es systematische Unterschiede in der räumlichen Verteilung der Subpopulationen gibt. Betrachte die Teststatistik

$$T = \sup_{B \text{ konvexes Gebiet}} D_B$$

mit

$$D_B = \frac{\text{Anzahl Einheiten der Subpopulation 1 in B}}{\text{Anzahl aller Einheiten von Subpopulation 1}} - \frac{\text{Anzahl Einheiten der Subpopulation 2 in B}}{\text{Anzahl aller Einheiten von Subpopulation 2}},$$

bzw., wenn man ein konvexes Gebiet durch den zugehörigen Begriffsinhalt als Schnitt von Halbräumen charakterisiert, in anderer Notation:

$$T = \sup_{(A,B) \in \mathfrak{B}((G,M,I))} D_{(A,B)}.$$

Dann ist $D_{(A,B)} \in [-1,1]$. Wir hatten den Extremfall betrachtet, dass alle Teilmengen von G Begriffsumfänge sind. In diesem Extremfall wird das Supremum für den Umfang A, der genau alle Einheiten aus Subpopulation 1 und keine Einheit aus Subpopulation 2 enthält, angenommen und ist unabhängig von der Verteilung der Subpopulationen immer konstant gleich 1. Dies bedeutet, dass T keine sinnvolle Teststatistik ist, in diesem Fall ist das Mengensystem der Begriffsumfänge aus statistischer Sicht einfach zu groß. (Vergleiche auch die späteren Abschnitte zur Extremaltheorie von Begriffsverbänden, zur Vapnik-Chervonenkis-Theorie und zur Regularisierung.) Daher die allgemeine Frage (die auch später noch weiter verfolgt wird): Wie kann man einen Begriffsverband kleiner machen? In der konkreten Situation hier kann man sicherlich viele Ideen haben:

i) Im Sinne eines Verständnisses von $\mathbb{R}^2$ als als eines metrischen Raums: Man könnte nicht beliebige konvexe Mengen, sondern beispielsweise nur metrisch einfach charakterisierbare Mengen[9] wie beispielsweise kreisförmige Mengen betrachten.

[8] Natürlich würde man wohl die beobachteten konvexen Mengen direkt in den Kontext mit aufnehmen, so dass man wieder bei Gegenstandsbegriffen landete. Ein anderes Beispiel, wo nicht direkt Gegenstandsbegriffe betrachtet werden, ist in der folgenden Anwendung gegeben. (Natürlich könnte man auch da die Nicht-Gegenstandsbegriffe nachträglich in den Kontext mit aufnehmen und sie so zu Gegenstandsbegriffen machen.)

[9] Man beachte aber, dass auch konvexe Mengen rein metrisch charakterisierbar sind via $y \in \text{co}\,\{x,z\} \iff d(x,z) = d(x,y) + d(y,z)$.

ii) Wenn man nicht direkt räumliche Daten hat, sondern zweidimensionale Daten, wo beide Komponenten inkommensurabel sind, dann könnte man im Sinne einer Forderung nach affiner Invarianz/Äquivarianz das Mengensystem aller ellipsenförmigen Mengen betrachten.

iii) Man könnte auch die Menge aller Begriffe, die von maximal k-elementigen Gegenstandsmengen erzeugt werden, also Begriffe der Form (A'', A') mit $|A| \leq k$ anschauen. Wenn k klein genug ist, dann würde das zugehörige System von formalen Begriffen auch klein genug sein.

iv) Alternativ könnte auch das System aller Begriffe, die von maximal k-elementigen Merkmalsmengen erzeugt werden, betrachten.

v) Alle obigen Mengensysteme sind keine Hüllensysteme und die Sprache der formalen Begriffsanalyse scheint deshalb für diese Lösungen wenig bereit stellen zu können.

Eine im Sinne der formalen Begriffsanalyse gangbare Möglichkeit wäre die folgende: Betrachte nicht den gesamten Kontext (G, M, I), sondern einen kleineren Kontext $(\tilde{G}, M, \tilde{I})$ mit $\tilde{G} \subsetneq G$ und $\tilde{I} = I \cap \tilde{G} \times M$. Berechne dann

$$\sup_{(A,B)\in\mathfrak{B}((\tilde{G},M,\tilde{I}))} D_{(A,B)}.$$

Achtung: $D_{(A,B)}$ wird im Gesamtkontext (G, M, I) und nicht im kleineren Kontext $(\tilde{G}, M, \tilde{I})$ berechnet, d.h, für das Zählen der Häufigkeiten werden alle Gegenstände aus G, die alle Eigenschaften aus B besitzen, gezählt. Außerdem beachte man, dass die Begriffsinhalte von $\mathfrak{B}((\tilde{G}, M, \tilde{I}))$ auch Begriffsinhalte von $\mathfrak{B}((G, M, I))$ sind. Die schwierige Frage ist nun jedoch noch, wie man $\tilde{G}$ geeignet und insbesondere nicht willkürlich wählen sollte? Eine Möglichkeit, die im Prinzip für beliebige Situationen anwendbar ist, ist die folgende: Ziehe zunächst zufällig (also in gewissem Sinne willkürlich) eine k-elementige Teilmenge $\tilde{G} \subseteq G$ (wobei jeder Gegenstand mit gleicher Wahrscheinlichkeit gezogen werde). Für jedes $\tilde{G}$ erhält man dann einen formalen Begriff $(A^*_{\tilde{G}}, B^*_{\tilde{G}})$, der $\sup_{(A,B)\in\mathfrak{B}((\tilde{G},M,\tilde{I}))} D_{(A,B)}$ maximiert. Jeder Begriffsinhalt $B^*_{\tilde{G}}$ gehört zu einem formalen Begriff im ursprünglichen Kontext (G, M, I), wir haben also eine Zufallsvariable mit Werten in einem vollständigen Verband und können das Quantilkonzept anwenden. (Man beachte, dass das Ziehen von $\tilde{G}$ willkürlich war, die Betrachtung der Verteilung der zugehörigen Begriffe, wenn $\tilde{G}$ über alle möglichen k-elementigen Teilmengen von G variiert, kuriert diese Willkür.) Insbesondere können wir den zentralsten formalen Begriff[10](bzw. die zentralsten Begriffe), zu dieser formal-begriffswertigen Zufallsvariable betrachten. Dieser ist dann eine repräsentativer Begriff, den wir zusammen mit der zugehörigen Teststatistik betrachten können. Konkret würde man also folgendermaßen vorgehen:

i) Für $l = 1, \ldots, N$ ziehe eine zufällige Gegenstandmenge $\tilde{G}_l \subseteq G$ der Größe k.

ii) Berechne $B^*_{\tilde{G}_l}$ und den zugehörigen Begriff $(B^*_{\tilde{G}_l}{}', B^*{}_{\tilde{G}_l})$.

iii) „Aggregiere“ die so erhaltenen formalen Begriffe durch Auswahl des zentralsten Begriffes (bzw. der zentralsten Begriffe).

iv) Dieser Begriff (A, B) zusammen mit dem zugehörigen Wert $D_{(A,B)}$ ist dann für einen „regularisierten“ Test verwendbar.

[10]Damit ist derjenige Begriff gemeint, der zum Quantil mit dem kleinstmöglichen α gehört.

Bemerkungen:

1. Der erhaltene Begriff und die erhaltene Teststatistik sind dann direkt im **urprünglichen** Kontext interpretierbar.

2. Die Teststatistik ist dann in dem Sinne regularisiert, dass man Begriffe im Ursprungskontext betrachtet, die derart sind, dass die Begriffsinhalte B von maximal k-elementigen Gegenstandsmengen A via $B = A'$ erzeugt werden. Es werden also maximal $\binom{|G|}{k}$ Begriffe betrachtet, was vergleichbar mit einer V.C.-Dimension k ist, vergleiche später. (Man beachte aber, dass der am Ende betrachtete Begriff nicht derjenige ist, für den $D_{(A,B)}$ maximal ist, sondern dass es sich um einen repräsentativen Begriff unter einer Verteilung von Begriffen handelt.)

3. Durch geeignete Wahl von k kann die Stärke der Regularisierung gesteuert werden.

4. Für $k = |G|$ erhält man die ursprüngliche nichtregularisierte Statistik.

5. Das Quantilkonzept kann auch genutzt werden, um die Streuung der zufällig gezogenen Begriffe deskriptiv zu analysieren.

6. Eine mögliche Modifikation des Vorgehens bestünde darin, am Ende nicht über alle erhaltenen Begriffe $(B^*_{\tilde{G}_l}{}', B^*_{\tilde{G}_l})$ zu aggregieren, sondern nur über diejenigen Begriffe $(B^*_{\tilde{G}_l}{}', B^*_{\tilde{G}_l})$, für die der Wert $D_{(B^*_{\tilde{G}_l}{}', B^*_{\tilde{G}_l})}$ nicht kleiner ist als ein Threshold c. (Alternativ könnte man auch die Begriffe entsprechend der zugehörigen Werte $D_{(B^*_{\tilde{G}_l}{}', B^*_{\tilde{G}_l})}$ gewichten.)

Abschließend soll das Quantilkonzept noch kurz für den Fall der deskriptiven Analyse von Ranking-Daten illustriert werden:

Beispiel 21 (**words** Datensatz, siehe Fligner and Verducci [1986]). Insgesamt 98 Studierende wurden gebeten, die 5 Worte 'Thought' (1), 'Play' (2),'Theory'(3),'Dream'(4) und 'Attention' (5) entprechend der Stärke ihrer Assoziation zu dem Referenzwort 'Idea' zu ranken. Tabelle 4 gibt die Daten wieder. Dabei sind die Rankings in sogenannter Rangnotation gegeben, z.B. der Vektor $(1, 3, 4, 5, 2)$ in der ersten Spalte und der ersten Zeile bedeutet, dass das zweite Wort 'Play' den Rang 3, bekommt, wobei Rang 1 für die geringste Assoziation und Rang 5 für die größte Asssoziation zum Referenzwort 'Idea' steht.

Ranking	Frequency	Ranking	Frequency	Ranking	Frequency
(1 3 4 5 2)	1	(4 2 3 5 1)	2	(5 1 4 2 3)	6
(1 4 2 3 5)	1	(4 3 5 2 1)	1	(5 1 4 3 2)	33
(3 2 5 4 1)	2	(5 1 2 4 3)	5	(5 2 3 4 1)	8
(4 1 2 5 3)	1	(5 1 3 2 4)	2	(5 2 4 1 3)	1
(4 1 5 3 2)	5	(5 1 3 4 2)	18	(5 2 4 3 1)	12

Tabelle 4: alle 98 Rankings des **words** Datensatzes.

Bei begrifflicher Skalierung wie im **wisdom of the crowd** Beispiel ergeben sich 156 formale Begriffe.

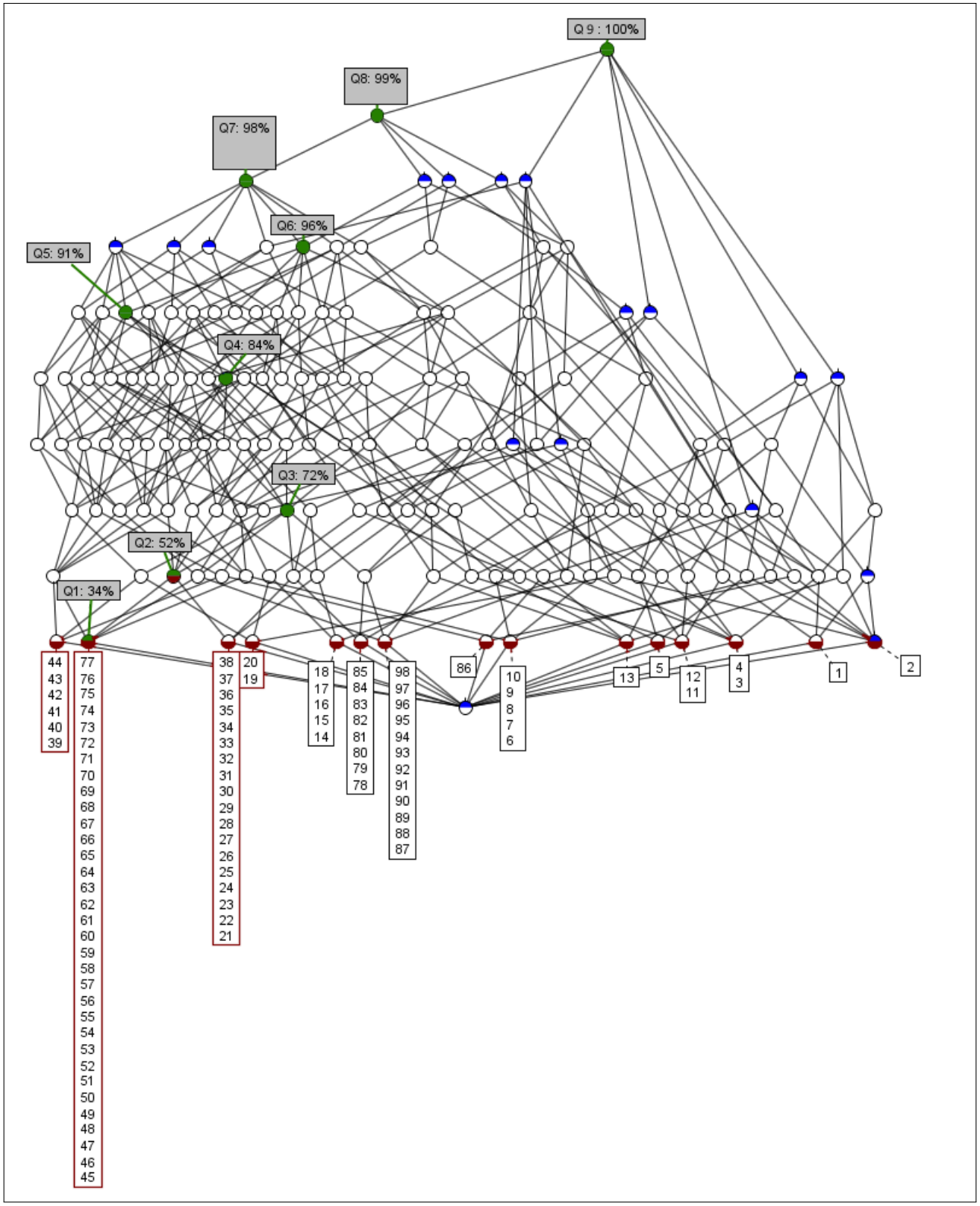


Abbildung 15: Der Begriffsverband des Kontextes zum **words** Datensatz mit 9 eingezeichneten Quantilen $Q_1, \ldots, Q_9$.

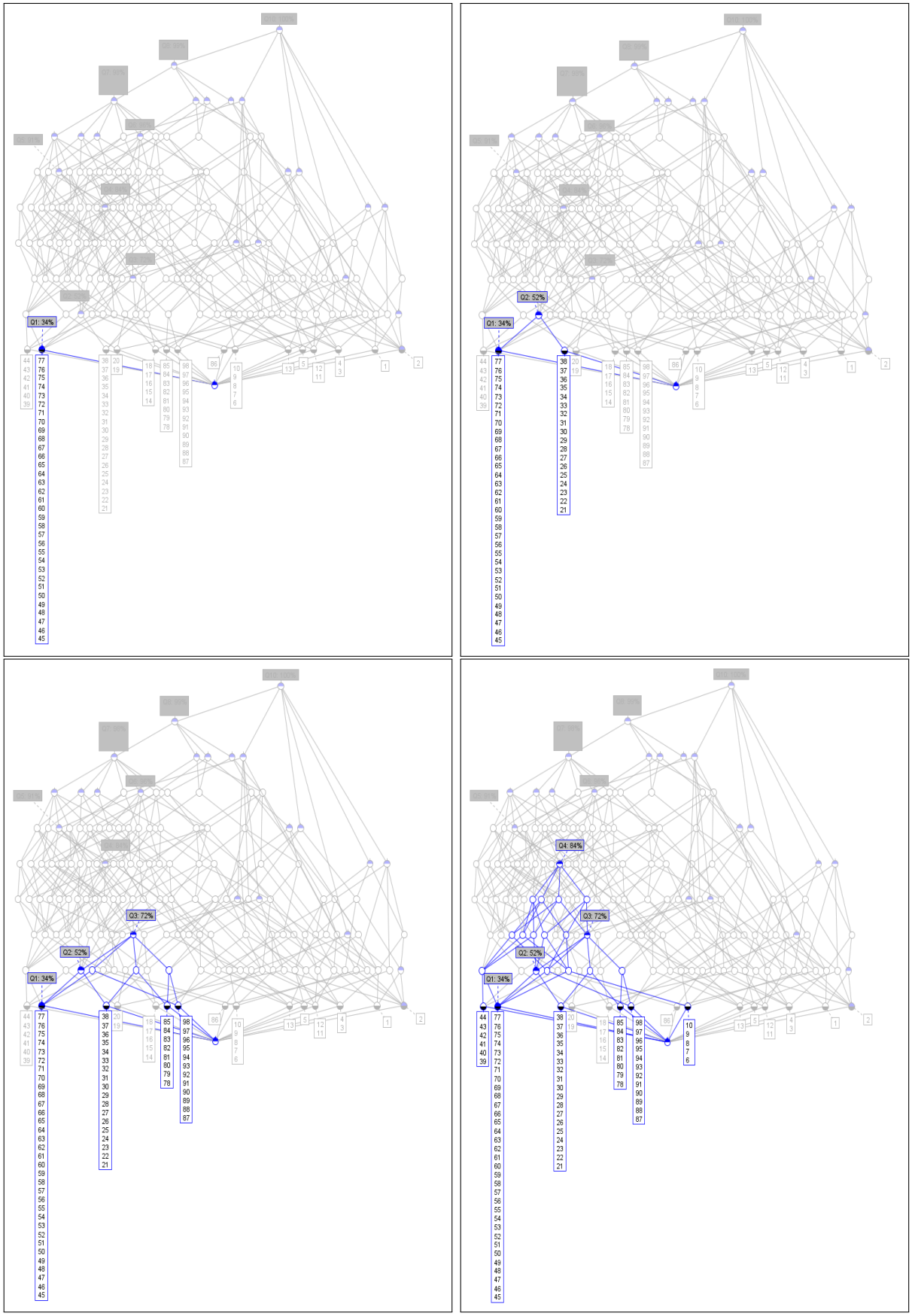


Abbildung 16: Das 34%-, 52%-, 72%- und das 84%-Quantil für den **words** Datensatz.

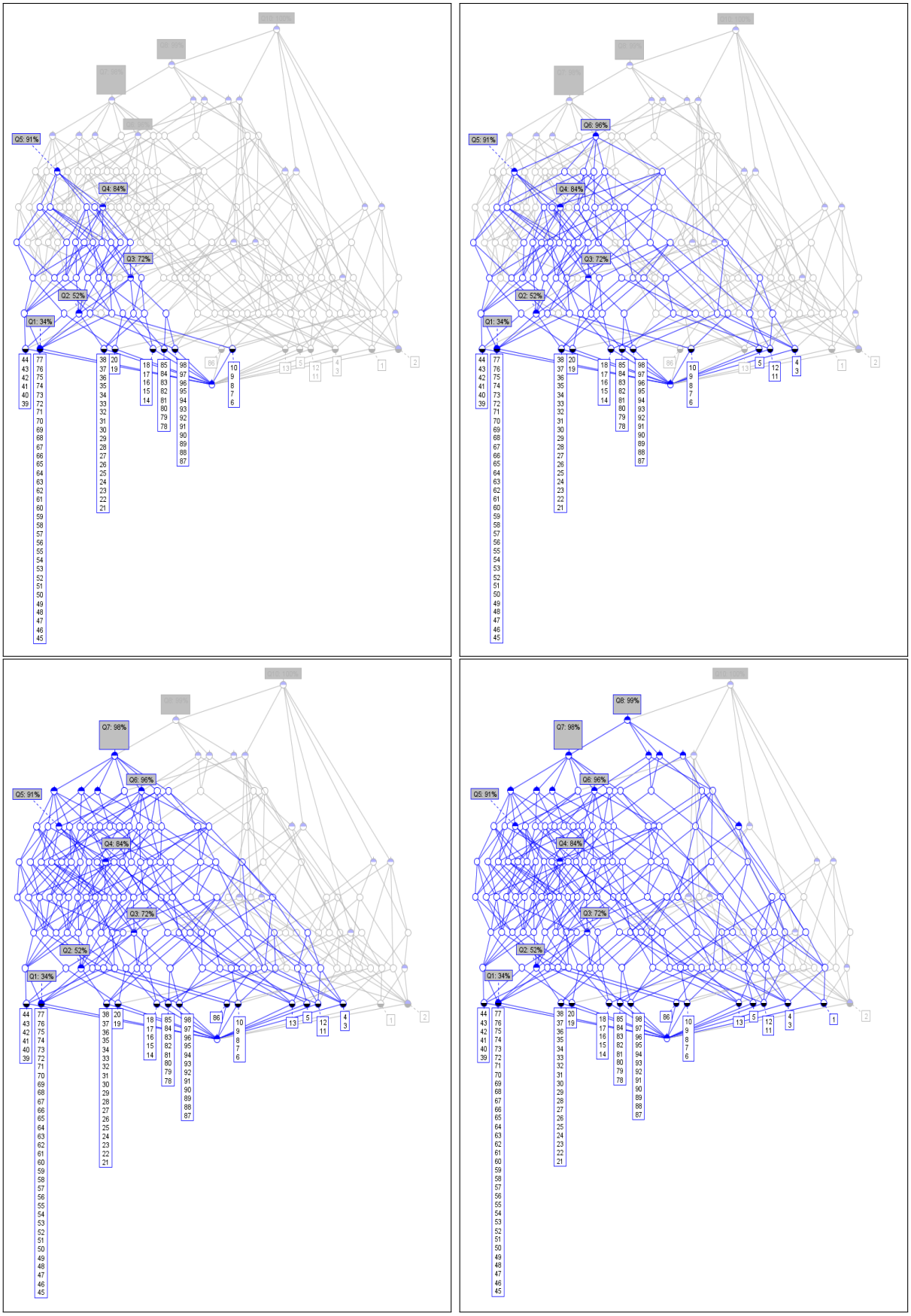


Abbildung 17: Das 91%-, 96%-, 98%- und das 99% Quantil für den **words** Datensatz.

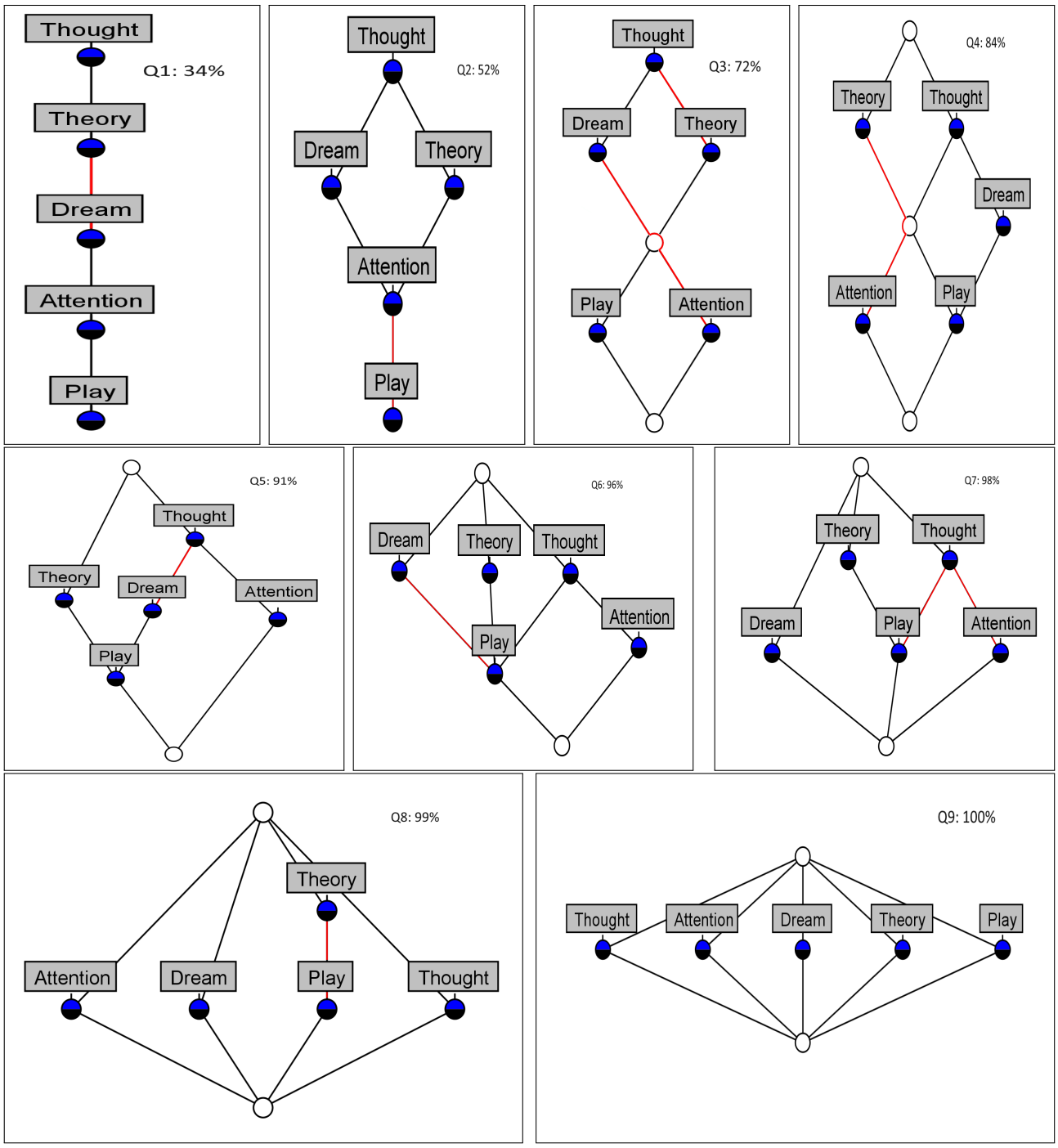


Abbildung 18: Die Begriffsinhalte der 9 Quantile $Q_1, \ldots, Q_9$.

Beispiel 22. Für das **wisdom of the crowd** Beispiel ist die Berechnung des gesamten Begriffsverbands sehr aufwendig aber immer noch möglich, es gibt hier insgesamt 10.452.368 Begriffe, was eine graphische Darstellung wenig sinnvoll erscheinen lässt. Demgegenüber ist das Bestimmen von Quantilen nach unseren Überlegungen natürlich immer noch möglich, es sind lediglich $44 \cdot 43 = 1892$ Gegenstandsbegriffe zu berechnen und anschließend die Quantile zu bilden. Insgesamt gibt es hier 4 Quantile. Die folgenden Graphiken zeigen jeweils die Begriffsinhalte der entsprechenden Quantile, also diejenigen Paare von Präsidenten, bezüglich derer sich die Personen des Umfangs des Quantils einig sind, welcher Präsident vorher bzw. nachher präsidiert hat. Da jede Person eine (totale) Ordnung angegeben hat, sind die Begriffsinhalte als Schnitt von Ordnungen immer noch partielle Ordnungen. Diese werden hier nicht direkt über den Hassegraph, sondern durch dass Hassediagramm des Begriffsverbands des Kontexts $(G, M, <)$ abgebildet, wobei $G = M$ die Menge der Präsidenten ist und $g < m$ gilt, falls alle Personen des Quantils einstimmig Präsident g zeitlich nach Präsident m einordnen. Die grauen labels in den Graphiken bezeichnen die Merkmale und die weißen labels bezeichnen die Gegenstände. Ein Präsident g wird einstimmig nach Präsident m eingeschätzt, falls der Begriff mit dem weißen label g ein Unterbegriff des Begriffs mit dem grauen lable m ist. Beispielsweise für das 7.7%-Quantil sieht man, dass Thomas Jefferson und John Adams zeitlich nach George Washington eingschätzt werden. Der Grund dafür, die Inhalte hier nicht über den Hassegraphen zur Relation $\leq$, sondern zur Relation $<$ darzustellen, liegt darin, dass die Diagramme dann etwas übersichtlicher sind, da die beobachteten Relationen relativ nahe an einer sogenannten Intervallordnung sind, näheres dazu sowie eine genauere deskriptive Analyse findet man in Schollmeyer [2017].

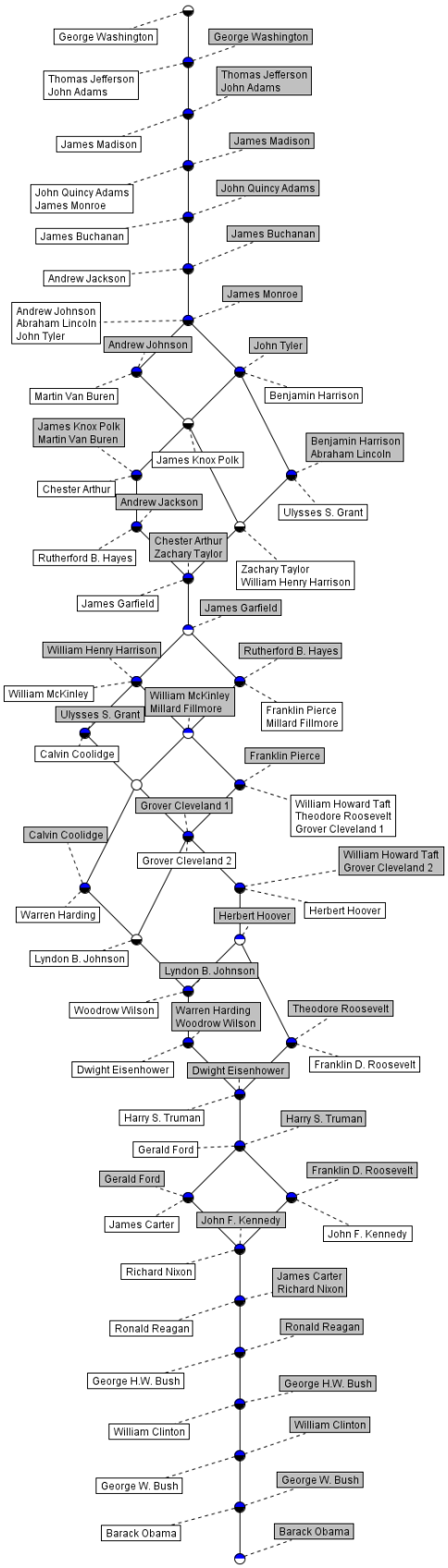


Abbildung 19: Das 7.7% Quantil für den US-Präsidenten Datensatz, repräsentiert über den formalen Begriffsverband des strikten Teils des Schnitts aller Rankings, die Gegenstände des 7.7% Quantils sind.

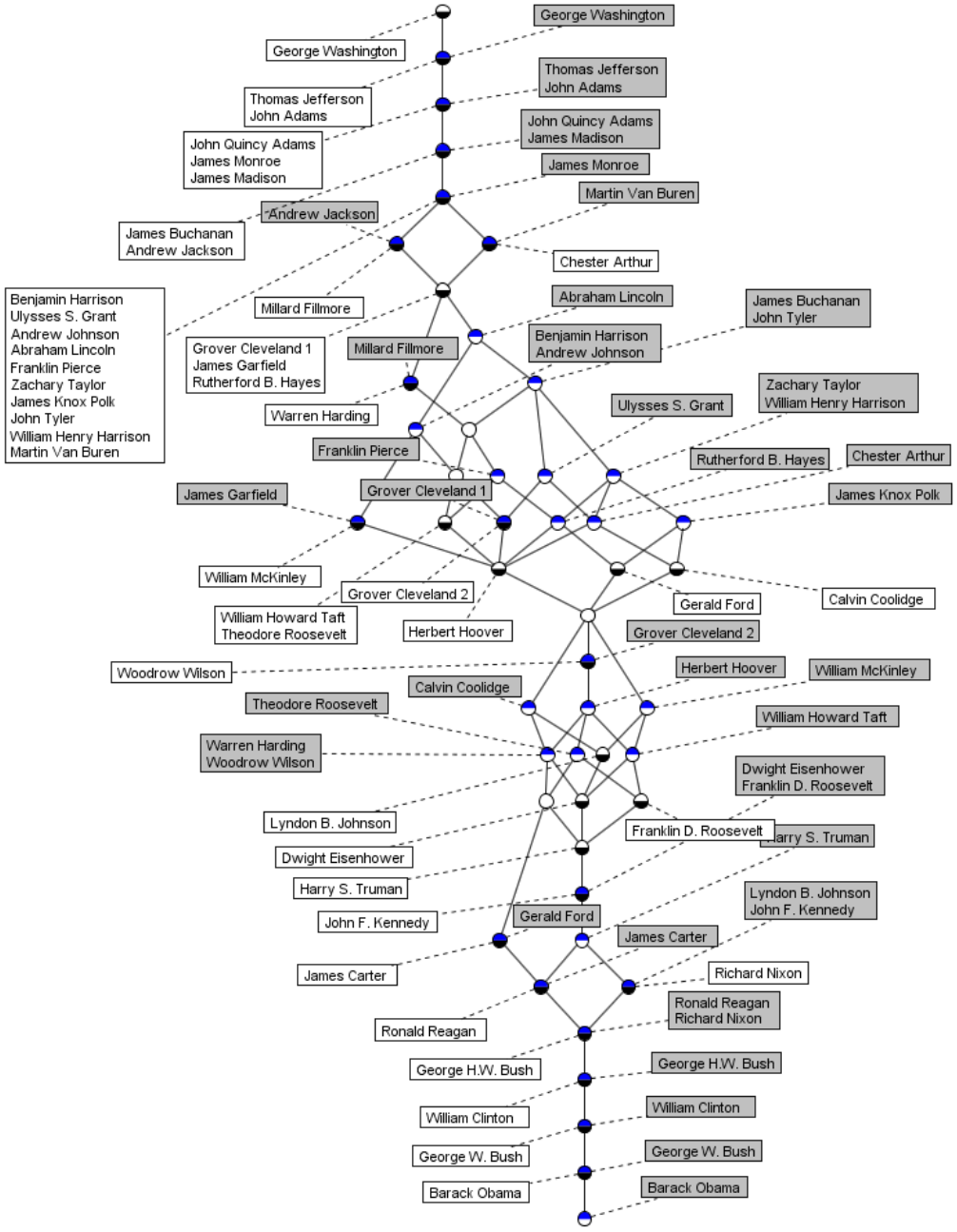


Abbildung 20: Das 27% Quantil für den US-Präsidenten Datensatz, repräsentiert über den formalen Begriffsverband des strikten Teils des Schnitts aller Rankings, die Gegenstände des 27% Quantils sind.

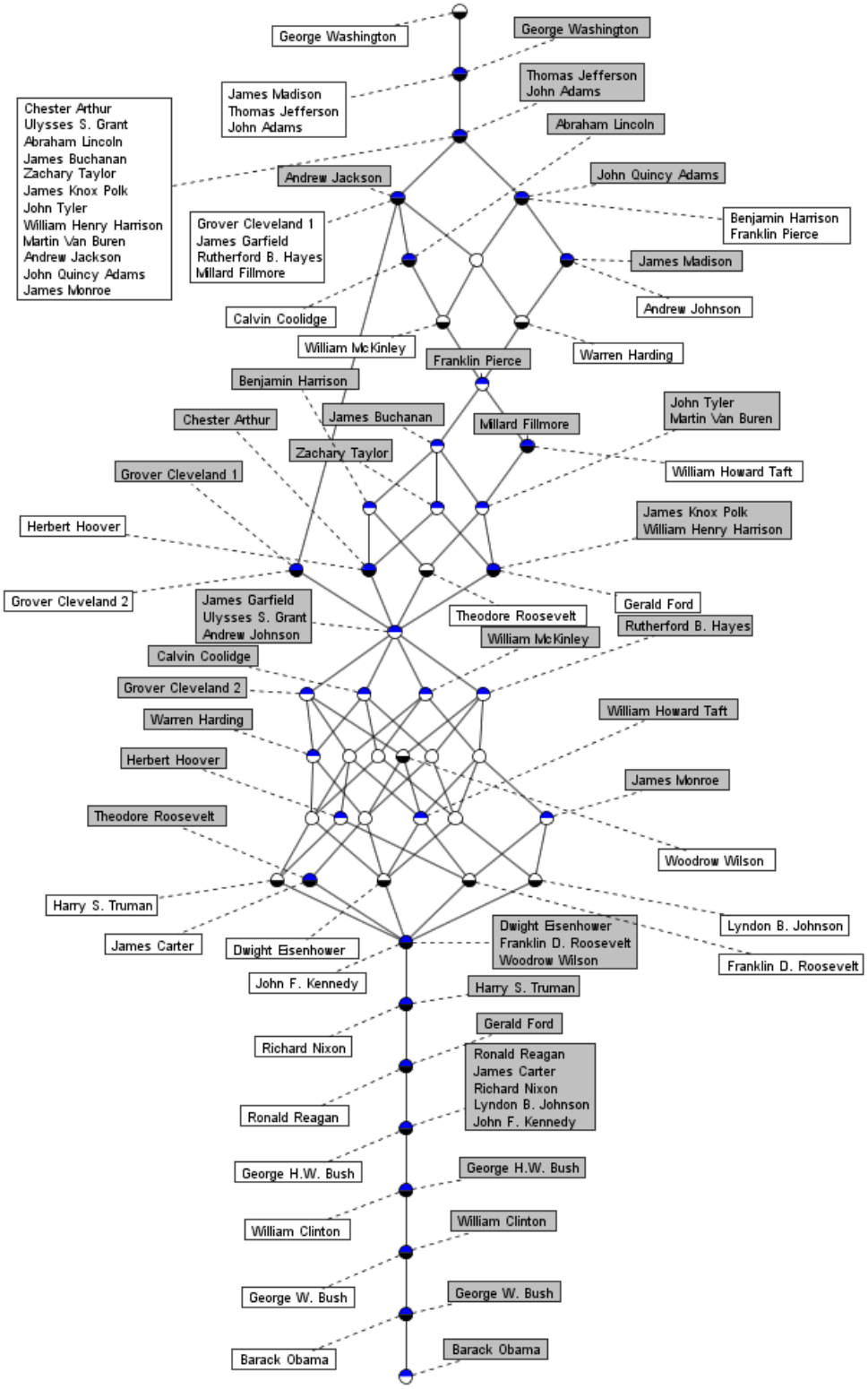


Abbildung 21: Das 42% Quantil für den US-Präsidenten Datensatz, repräsentiert über den formalen Begriffsverband des strikten Teils des Schnitts aller Rankings, die Gegenstände des 42% Quantils sind.

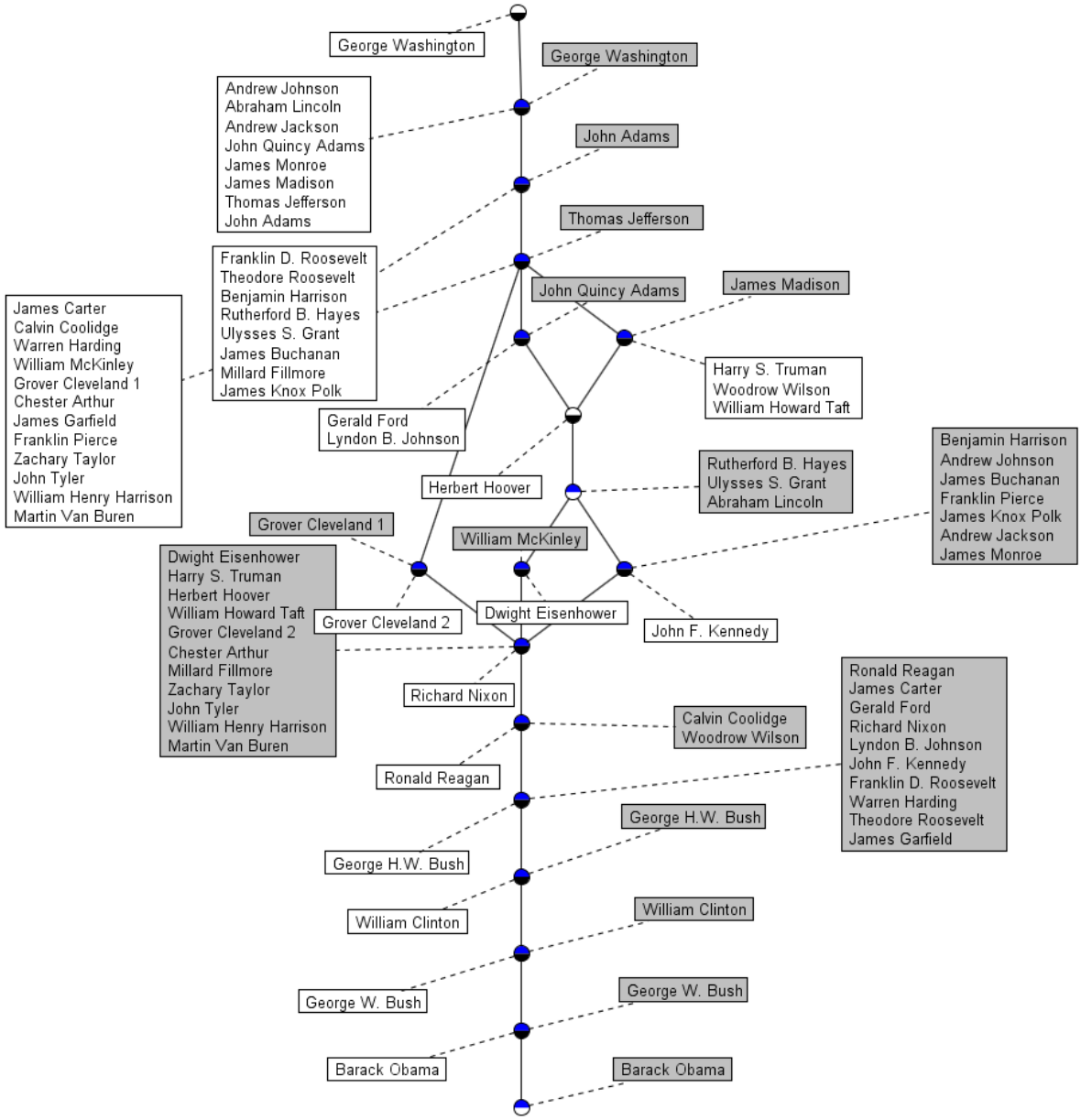


Abbildung 22: Das 100% Quantil für den US-Präsidenten Datensatz, repräsentiert über den formalen Begriffsverband des strikten Teils des Schnitts aller 26 Rankings.

3 Stochastische Dominanz

Fragestellung: Gegeben zwei Zufallsvariablen $X, Y : \Omega \longrightarrow V$ mit Werten in einer geordneten Menge $(V, \leq)$, wie kann man „sinnvoll formalisieren“, dass X stochastisch kleinergleich Y ist? Wir wollen also die Relation $\leq$ zwischen Elementen aus V auf Zufallsvariablen mit Werten in V übertragen, um in gewissen Situationen davon sprechen zu können, dass eine zufällige Größe X stochastisch kleinergleich einer zufälligen Größe Y ist. Dies könnte man vergleichen mit dem Ansinnen bei der Definition des Erwartungswertes einer Zufallsvariablen, den Wert einer reellen Zahl auf den Wert einer zufälligen reellen Zahl zu übertragen. Überlegungen:

a) Falls $\forall \omega \in \Omega : X(\omega) \leq Y(\omega)$, so ist X mit Sicherheit kleinergleich Y.

b) Falls $P(\{\omega \in \Omega \mid X(\omega) \leq Y(\omega)\}) = 1$, dann ist X fast sicher kleinergleich Y.

c) Wir wollen uns nun gedanklich vom unterliegenden Grundraum Ω lösen und nur die Verteilung von X bzw. Y heranziehen. Dazu betrachten wir:

   i) Einfachste Situation: $X \equiv a$, $Y \equiv b$ mit $a, b \in V$ und $a \leq b$. Dann kann man wohl sagen, dass X stochastisch kleinergleich Y ist. (In dieser Situation ist X sicher kleinergleich Y.) Dass X stochastisch kleinergleich Y ist, werden wir im Folgenden mit $X \leq_{SD} Y$ notieren.

   ii) Auch für $P(X = a) = 1$ und $P(Y = b) = 1$ mit $a \leq b$ würde man wohl sagen, dass $X \leq_{SD} Y$ gilt.

   iii) Nächst schwierigere Situation: $a \leq b$ und $X = \begin{cases} b & \text{mit Wahrscheinlichkeit } p \\ a & \text{mit Wahrscheinlichkeit } 1-p \end{cases}$ sowie $Y = \begin{cases} b & \text{mit Wahrscheinlichkeit } q \\ a & \text{mit Wahrscheinlichkeit } 1-q \end{cases}$. Dann ist naheliegend zu definieren:
   $X \leq_{SD} Y :\iff p \leq q$. In Worten etwa: $X \leq_{SD} Y$ falls die Wahrscheinlichkeit dafür, dass X große Werte (hier konkret den Wert b) annimmt, kleinergleich ist als die Wahrscheinlichkeit dafür, dass Y große Werte annimmt.

   iv) Noch schwierigere Situation: $a < b < c$ und $X(\omega) \in \{a, b, c\}$ sowie $Y(\omega) \in \{a, b, c\}$:
   $X \leq_{SD} Y \iff$ Die Wahrscheinlichkeit dafür, dass X große Werte annimmt ist kleinergleich der Wahrscheinlichkeit dafür, dass Y große Werte annimmt, unabhängig davon, wie man den Ausdruck „große Werte“ **sinnvoll** konkretisiert. In der konkreten Situation wäre eine sinnvolle Konkretisierung von „groß“ gegeben durch $\in \{b, c\}$ oder $\in \{a, b, c\}$ oder $\in \{c\}$ oder auch $\in \emptyset$. Nicht sinnvoll wäre wohl beispielsweise $\in \{a, b\}$, da ja c größer ist als a und b, und wenn selbst a schon als „groß“ bezeichnet wird, warum soll dann das größere c nicht als groß bezeichnet werden? Ganz allgemein könnte man die Forderung nach einer **sinnvollen** Konkretisierung des Ausdrucks „groß“ so formulieren bzw. formalisieren: Mit jedem Element $x \in V$, das als groß deklariert wird, sollte auch jedes $y \in V$ mit $y \geq x$ als groß deklariert werden. Also: jede **sinnvolle** Deklaration von Werten als „groß“ kann durch eine Oberhalbmenge $A \subseteq V$ beschrieben werden über

$$x \text{ ist groß} \iff x \in A.$$

   Die Eigenschaft einer Menge $A \subseteq V$, Oberhalbmenge zu sein, bedeutet dann genau, dass es sich (natürlich nur aus rein mathematischer Sicht) um eine sinnvolle Deklaration von Werten als „groß“ handelt.

Mit diesen Überlegungen sind wir jetzt in der Lage, das Konzept der stochastischen Dominanz zu definieren.

Definition 3.1 (Stochastische Dominanz erster Ordnung, Oberhalbmengenfassung) Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X, Y : \Omega \longrightarrow V$ Zufallsvariablen mit Werten in einer geordneten Menge $(V, \leq)$, die mit der σ-Algebra $\mathcal{A}'$ ausgestattet sei. Dann heißt X stochastisch kleinergleich Y (in Zeichen: $X \leq_{SD} Y$) bezogen auf stochastische Dominanz erster Ordnung, falls für jede $\mathcal{A}'$-messbare Oberhalbmenge S von V gilt:

$$P(X \in S) \leq P(Y \in S).$$

Bemerkung 3.1. Die Relation $\leq_{SD}$ ist eine Quasiordnung, also transitiv und reflexiv. Im Allgemeinen ist $\leq_{SD}$ nicht antisymmetrisch, beispielsweise für X und Y mit $X \neq Y$ aber $X \stackrel{d}{=} Y$ gilt gleichzeitig $X \leq_{SD} Y$ und $Y \leq_{SD} X$.

Bemerkung 3.2. Da man bei der stochastischen Dominanz lediglich auf das Bildmaß der unterliegenden Zufallsvariablen rekurriert, ist es auch möglich, Zufallsvariablen auf verschiedenen unterliegenden Wahrscheinlichkeitsräumen zu betrachten.

Bemerkung 3.3. Neben der stochastischen Dominanz gibt es noch das Konzept der statistischen Präferenz, das nicht nur auf die Bildmaße rekurriert, konkret definiert man

$$X \leq_{SP} Y :\iff P(X \leq Y) \geq 0.5.$$

Diese Relation ist reflexiv und total, aber nicht transitiv. Darüber hinaus gibt es noch Konzepte stochastischer Dominanz höherer Ordnung, die aber nicht von rein ordnungstheoretischer Natur sind und üblicherweise nur definiert werden für rellwertige Zufallsvariablen, $\mathbb{R}^k$-wertige Zufallsvariablen oder Zufallsvariablen mit Werten in einer Menge, die über eine Ordnungsstruktur hinaus noch mit irgendeiner Form von kardinaler Struktur ausgestattet ist.

Beispiel 23. Sei $(V, \leq) = (\mathbb{R}, \leq)$. Dann sind die Oberhalbmengen genau die Intervalle der Form $]c, \infty]$ bzw. $[c, \infty]$ mit $c \in \mathbb{R} \cup \{-\infty, \infty\}$. Für reellwertige Zufallsvariablen ist deshalb stochastische Dominanz über die zugehörigen Verteilungsfunktionen charakterisierbar qua

$$X \leq_{SD} Y \iff \forall c \in \mathbb{R} : \underbrace{F_X(c)}_{=P(X \leq c) = 1 - P(X > c)} \geq \underbrace{F_Y(c)}_{=P(Y \leq c) = 1 - PY > c}.$$

(Man beachte, dass gilt: $P(X \geq c) = P\left(\bigcap_{n \in \mathbb{N}} \{X > c - 1/n\}\right) = \lim_{n \to \infty} P(X > c - 1/n) = \lim_{n \to \infty} 1 - F_X(c - 1/n)$.) Diese Charakterisierung wird oft direkt als Definition für stochastische Dominanz erster Ordnung für reellwertige Zufallsvariablen verwendet.

Satz 3.2 (Charakterisierung stochastischer Dominanz erster Ordnung) Im Wesentlichen[11] sind die folgenden Aussagen äquivalent:

i) $\forall S \subseteq V$ messbare Oberhalbmenge $: P(X \in S) \leq P(Y \in S)$.

ii) $\exists$ Wahrscheinlichkeitsraum $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$, $\exists$ Zufallsvariablen $\tilde{X}, \tilde{Y} : \tilde{\Omega} \longrightarrow V$:

$$\tilde{X} \stackrel{d}{=} X, \tilde{Y} \stackrel{d}{=} Y, \tilde{P}(\tilde{X} \leq \tilde{Y}) = 1.$$

[11] Hier muss vorausgesetzt werden, dass $(V, \leq)$ ein polnischer Raum (also ein separabler, vollständig metrisierbarer topologischer Raum) ist, und dass die Menge $\{(x, y) \mid x, y \in V, x \leq y\}$ abgeschlossen in der Produkttopologie auf $V \times V$ ist.

iii) Für jede beschränkte, messbare isotone Funktion $u : V \longrightarrow \mathbb{R}$ gilt

$$\mathbb{E}(u \circ X) \leq \mathbb{E}(u \circ Y).$$

iv) Für jede lineare Erweiterung L von $\geq$ gilt: $\forall c \in V : P(XLc) \leq P(YLc)$.

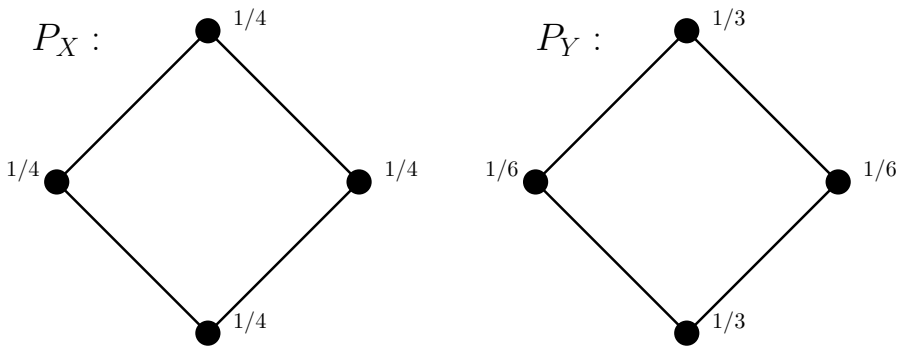
v) Man kann die Verteilung von Y aus der Verteilung von X durch einen Transport der Wahrscheinlichkeitstsmasse von X hin zu größeren Werten erhalten.[12]

Bemerkung 3.4. Im Allgemeinen sind die Bedingungen

i) $\forall S \subseteq V$ messbare Oberhalbmenge: $P(X \in S) \leq P(Y \in S)$ und

ĩ) $\forall c \in V$ mit $\uparrow c$ messbar: $P(X \in \uparrow c) \leq P(Y \in \uparrow c)$

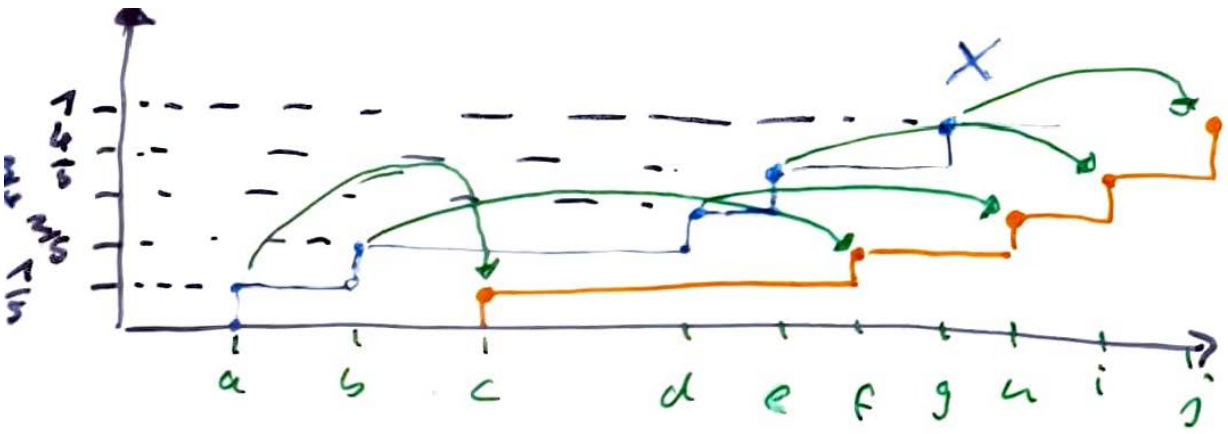
nicht äquivalent. Minimalbeispiel dazu:



Hier ist $P(X \geq c) \leq P(Y \geq c)$ für alle $c \in V$, aber

$$P\left(X \in \right) = 3/4 > 2/3 = P\left(Y \in \right).$$

Beispiel 24 (Ein einfaches Besipiel mit $(V, \leq)$ total geordnet zur Illustration obiger Charakterisierung). Es seien X und Y reellwertige, diskrete Zufallsvariablen mit Punktwahrscheinlichkeiten von jeweils $\frac{1}{n}$ mit z.B. $n = 5$.



[12]Dies ist natürlich nur eine sehr informelle Beschreibung, die nur äquivalent, ist, wenn P_X und P_Y als Wahrscheinlichkeitsmaße mit Dichte bezüglich eines gemeinsamen Maßes darstellbar sind.

- Charakterisierung über Verteilungsfunktion:

$$F_X(c) \geq F_Y(c) \text{ für alle } c \in \mathbb{R}.$$

- Für Charakterisierung über ii) betrachte

$$\tilde{\Omega} = \{\omega_1, \ldots, \omega_5\}$$
$$\tilde{\mathcal{A}} = 2^{\tilde{\Omega}}$$
$$\tilde{P}(\{\omega_i\}) = \frac{1}{5}, i = 1, \ldots, 5$$
$$\tilde{X} : \begin{cases} \omega_1 & \mapsto g \\ \omega_2 & \mapsto e \\ \omega_3 & \mapsto d \\ \omega_4 & \mapsto b \\ \omega_5 & \mapsto a \end{cases}$$
$$\tilde{Y} : \begin{cases} \omega_1 & \mapsto j \geq g = \tilde{X}(\omega_1) \\ \omega_2 & \mapsto i \geq e = \tilde{X}(\omega_2) \\ \omega_3 & \mapsto h \geq d = \tilde{X}(\omega_3) \\ \omega_4 & \mapsto f \geq b = \tilde{X}(\omega_4) \\ \omega_5 & \mapsto c \geq a = \tilde{X}(\omega_5) \end{cases}$$

  Damit: $\tilde{P}(\tilde{X} \leq \tilde{Y}) = 1$.

- Zur Charakterisierung über iii): Um hier $\mathbb{E}(u \circ X) \leq \mathbb{E}(u \circ Y)$ für jedes isotone, messbare und beschränkte u einzusehen, beachte man einfach, dass die Verteilungsfunktion von X immer oberhalb der Verteilungsfunktion von Y liegt. Damit liegt auch für jedes isotone u die Verteilungsfunktion von $u \circ X$ immer oberhalb der Verteilungsfunktion von $u \circ Y$. Mit der Darstellung des Erwartungswertes einer Zufallsvariablen Z über die Verteilungsfunktion F_Z als $\mathbb{E}(Z) = \int_0^\infty 1 - F_Z(t)\ dt - \int_{-\infty}^0 F_Z(t)\ dt$ (vergleiche beispielsweise Muldowney et al. [2012]) sieht man mit der Isotonie des Integrals sofort, dass dann $\mathbb{E}(u \circ X) \leq \mathbb{E}(u \circ Y)$ gilt.

- Zur Charakterisierung über iv): Hier ist $\mathbb{R}$ ja schon total geordnet. Allgemein gilt für $(V, \leq)$ nicht total geordnet und L eine beliebige totale Erweiterung von $\leq$, dass die Oberhalbmengen (V, L) auch Oberhalbmengen von $(V, \leq)$ sind. Außerdem kann man für jede Oberhalbmenge A von $(V, \leq)$ die Elemente von A untereinander beliebig (unter Respektierung von $\leq$) linear anordnen und die Elemente aus A^c alle unterhalb von allen Elementen von A beliebig (unter Respektierung von $\leq$) linear anordnen und erhält somit eine lineare Erweiterung L und eine Oberhalbmenge in (V, L), nämlich genau A. Dies bedeutet, dass man sich jede Oberhalbmenge von $(V, \leq)$ vorstellen kann als entstanden durch eine Linearisierung von $(V, \leq)$ und ein anschließendes Auswählen einer Oberhalbmenge bezüglich der linearen Erweiterung.

- Charakterisierung über v): Siehe grüne Pfeile in obiger Abbildung.

Achtung: So einfach ist es natürlich nur im total geordneten Fall. Beispielsweise der Massetransport vom größten Wert von X auf den größten Wert von Y, vom zweitgrößten Wert von

X auf den zweitgrößten Werte von Y usw. ist nicht einfach auf den partiell geordneten Fall übertragbar.

Im Folgenden werden wir nicht analytisch gegebene Zufallsvariablen X, Y betrachten, sondern wir werden den Fall betrachten, dass wir von X und Y nur Stichproben mit Realisierungen $x_1, \ldots, x_{n_x}$ bzw. $y_1, \ldots, y_{n_y}$ vorliegen haben. In dem Fall wollen wir von empirischer stochastischer Dominanz (in Zeichen $X \leq_{\hat{SD}} Y$) sprechen, falls für Zufallsvariablen mit Verteilungen, die exakt den empirischen Verteilungen in den Stichproben entsprechen, entsprechend stochastische Dominanz vorliegt. Dann kann man natürlich fragen, inwieweit von empirischer Dominanz in der Stichprobe auf wirklich vorliegende Dominanz geschlossen werden kann:

Zur statistischen Inferenz

Vergleiche die Grundidee beim Kolmogorov-Smirnov-Test: Da wird die wahre unbekannte Verteilungsfunktion F_X einer Zufallsvariablen X durch die empirische Verteilungsfunktion F_n einer i.i.d.-Stichprobe $x_1, \ldots, x_n$ geschätzt. (Die empirische Verteilungsfunktion ist die Maximum-Likelihood-Schätzung im völlig nichtparametrischen Setting.) Der Hauptsatz der Statistik sichert, dass die empirische Verteilungsfunktion gleichmäßig gegen die wahre unbekannte Verteilungsfunktion konvergiert:

$$\sup_{x \in \mathbb{R}} | \underbrace{F_n(x)}_{=\hat{P}_X(]-\infty,x])} - \underbrace{F_X(x)}_{=P_X(]-\infty,x])} | \overset{n\to\infty}{\longrightarrow} 0 \text{ P-fast sicher .}$$

Für

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_X$$
$$Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} F_Y$$

und den Einstichproben-Test

$$H_0 : F_X = F_0 \quad \text{vs} \quad H_1 : F_X \neq F_0$$

kann die Teststatistik

$$T := \sup_{x \in \mathbb{R}} | \underbrace{\hat{F}_n^X(x)}_{:=\frac{|\{i|x_i \leq x\}|}{n}} - F_0(x)|$$

verwendet werden. Analog kann für den Zweistichproben-Test

$$H_0 : F_X = F_Y \quad \text{vs} \quad H_1 : F_X \neq F_Y$$

die Teststatistik

$$T := \sup_{x \in \mathbb{R}} | \underbrace{\hat{F}_n^X(x)}_{:=\frac{|\{i|x_i \leq x\}|}{n}} - \underbrace{\hat{F}_m^Y(x)}_{:=\frac{|\{i|y_i \leq x\}|}{m}} |$$

verwendet werden. In beiden Fällen gilt unter H_0, dass $T \overset{m,n\to\infty}{\longrightarrow} 0$ fast sicher und unter H_1 gilt $T \overset{m,n\to\infty}{\longrightarrow} c > 0$ fast sicher, d.h., die Teststatistik T führt jeweils zu einem konsistenten Test.

Für das Testen auf stochastische Dominanz könnte man meinen, ähnlich vorgehen zu können, indem man sich überlegte:

$$\begin{aligned} X \leq_{SD} Y &\iff \forall S \in \mathfrak{O}((V,\leq)) \text{ messbar } : P(X \in S) \leq P(Y \in S) \\ &\iff \forall S \in \mathfrak{O}((V,\leq)) \text{ messbar } : P(X \in S) - P(Y \in S) \leq 0 \\ &\iff \sup_{\substack{S \in \mathfrak{O}((V,\leq)) \\ \text{messbar}}} P_X(S) - P_Y(S) \leq 0, \end{aligned}$$

d.h., eine naheliegende Teststatistik wäre hier

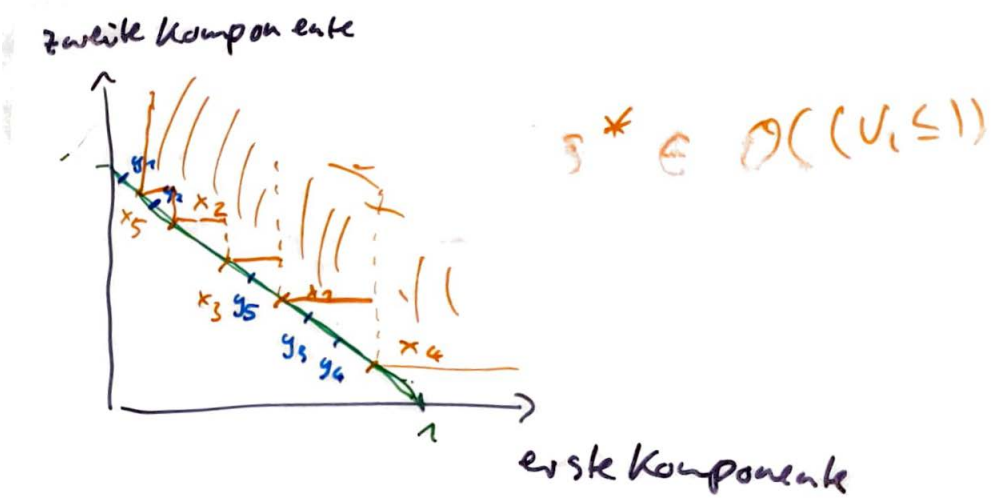
$$T := \sup_{\substack{S \in \mathfrak{O}((V,\leq)) \\ \text{messbar}}} \hat{P}_n^X(S) - \hat{P}_m^Y(S)$$

mit $\hat{P}_n^X(S) = \frac{|\{i \in \{1,\ldots,n\} | x_i \in S\}|}{n}$ und $\hat{P}_m^Y(S) = \frac{|\{i \in \{1,\ldots,m\} | y_i \in S\}|}{m}$. Die Frage wäre nun zuächst auch, wie sich die Statistik T aus statistischer Sicht verhält, bzw. konkret, ob unter $X \leq_{SD} Y$ die Statistik T für n, m gegen unendlich gegen 0 konvergiert. Dem ist im Allgemeinen nicht so, betrachte das folgende Beispiel:

Beispiel 25. Sei $(V,\leq) = (\mathbb{R}^2,\leq)$ mit $x \leq y \iff x_1 \leq y_1 \;\&\; x_2 \leq y_2$ und $(\Omega,\mathcal{A},P) = ([0,1],\mathfrak{B}([0,1]),\lambda_{|[0,1]})$, sowie

$$X : \Omega \longrightarrow \mathbb{R}^2 : \omega \mapsto \begin{pmatrix} \omega \\ 1-\omega \end{pmatrix},$$

$Y \overset{d}{=} X$ und $X_1,\ldots,X_n,Y_1,\ldots,Y_n \overset{i.i.d.}{=} X$ mit Realisierungen $x_1,\ldots,x_n,y_1,\ldots,y_n$. Dann sind alle Realisierungen fast sicher paarweise verschieden.



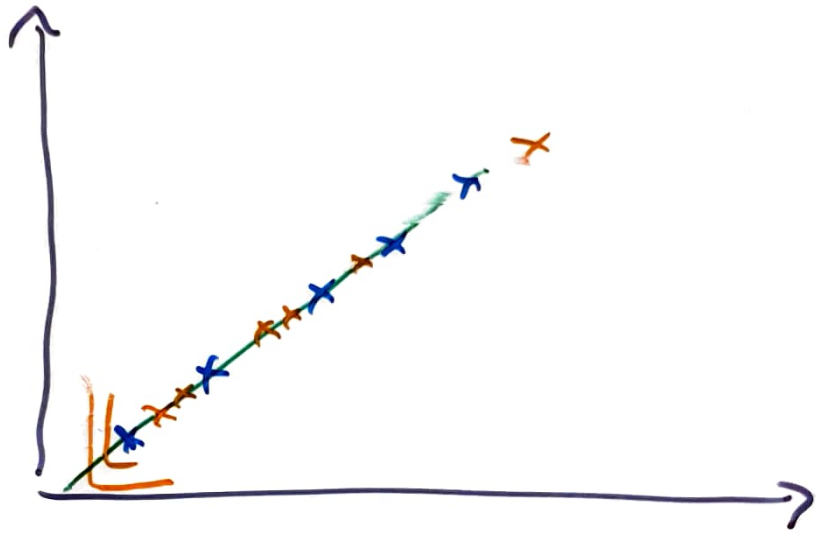
Nach obiger Skizze gilt deshalb für die Teststatistik T fast sicher:

$$T = \sup_{\substack{S \in \mathfrak{O}((\mathbb{R}^2,\leq)) \\ \text{messbar}}} \hat{P}_n^X(S) - \hat{P}_n^Y(S) \geq \hat{P}_n^X(S^*) - \hat{P}_n^Y(S^*) = 1,$$

also $P(T = 1) = 1$, d.h., in dieser extremen Situation ist die Teststatistik T unbrauchbar.

Frage: Warum betrachten wir dann Obiges überhaupt?

a) Obiges Beispiel ist nur ein Extremfall, die Situation könnte auch viel gutartiger sein, beispielsweise so:



Außerdem: Um einzuschätzen, wie arg die Situation ist, genügt es, die Menge $\{x_1, \ldots, x_n, y_1, \ldots, y_n\}$ zu kennen, man muss nicht die Zugehörigkeit der Realsierungen zu den Stichproben X bzw. Y kennen. Dies bedeutet insbesondere, dass beispielsweise ein Permutationstest auch nach einer Analyse der Argheit der Situation (und eventueller „Regularsierung“, vgl. später) noch valide bleibt, sofern die Zugehörigkeit der Realsisierungen zu den Stichproben X bzw. Y vorher nirgendwo einbezogen wurde.

b) Wenn es wirklich arg ist: „Regularisierung“ (Betrachte nicht das Mengensystem aller Oberhalbmengen, sondern ein kleineres Mengensystem, siehe später.)

Zur Berechnung der Supremumstatstik T

Notation: Es sei $(V, \leq) = (\{v_1, \ldots, v_k\}, \leq)$ eine endliche[13] geordnete Menge, die mit der Potenzmenge $\mathcal{A}' = 2^V$ als σ-Algebra ausgestattet sei. Sei weiter $x = (x_1, \ldots, x_{n_x})$ ein i.i.d.-Sample einer Zufallsvariablen X und $y = (y_1, \ldots, y_{n_y})$ ein i.i.d.-Sample einer Zufallsvariablen Y. Definiere $w^x = (w_1^x, \ldots, w_k^x)$ und $w^y = (w_1^y, \ldots, w_k^y)$ mit

$$w_i^x = \frac{\text{Anzahl Fälle in Stichprobe } x\text{, in denen der Wert } v_i \text{ beobachtet wurde}}{n_x}$$
$$w_i^y = \frac{\text{Anzahl Fälle in Stichprobe } y\text{, in denen der Wert } v_i \text{ beobachtet wurde}}{n_y}.$$

[13] Dies ist hier keine Einschränkung, für jede unendliche geordnete Menge $(V, \leq)$ kann man für das Feststellen empirischer Dominanz einfach die endliche Menge V_{obs} aller beobachteten Elemente aus V betrachten, denn es gilt: Ist A eine Oberhalbmenge in $(V, \leq)$, so ist $A \cap V_{obs}$ eine Oberhalbmenge in $(V_{obs}, \leq \cap V_{obs} \times V_{obs})$ und es gilt $\hat{P}_n^X(A) = \hat{P}_n^X(A \cap V_{obs})$ sowie $\hat{P}_m^Y(A) = \hat{P}_m^Y(A \cap V_{obs})$. Andersrum ist für jede Oberhalbmenge A von $(V_{obs}, \leq \cap V_{obs} \times V_{obs})$ die Menge $\tilde{A} := A \cup \{x \in V \backslash V_{obs} \mid \exists a \in A : x \geq a\}$ eine Oberhalbmenge in $(V, \leq)$ mit $\hat{P}_n^X(\tilde{A}) = \hat{P}_n^X(A)$ und $\hat{P}_m^Y(\tilde{A}) = \hat{P}_m^Y(A)$. **Achtung:** So schön ist es nur für den Zweistichprobentest, für den Einstichprobentest kommt man im Allgemeinen nicht mit dem Betrachten nur der beobachteten Werte aus V aus. Schon im Fall, dass man nicht das System aller Oberhalbmengen, sondern nur alle Hauptfilter (dies sind Mengen der Form $\uparrow c$) betrachtet, ist diese Unschönheit gegeben, vergleiche dazu beispielsweise die Unterschiede, die sich zwischen bivariaten Verteilungstests nach [Fasano and Franceschini, 1987] und nach [Peacock, 1983] ergeben, sowie allgemein die Implikationen die sich aus dieser Unschönheit im Kontext von Inferenz ergeben im Sinne einer Unterscheidung von *induktiver* und *transduktiver* Inferenz, siehe [Vapnik, 2006, Kapitel 3.1].

Die Menge aller Oberhalbmengen von $(V, \leq)$ sei mit $\mathfrak{O}((V, \leq))$ bezeichnet. Für $A \subseteq V$ definiere den charakteristischen Vektor $s^A \in \{0,1\}^k$ als

$$s_i^A = \begin{cases} 1 & \text{falls } v_i \in A \\ 0 & \text{sonst} \end{cases}.$$

Schließlich sei

$$B_V = \{s \in \{0,1\} \mid \forall (v_i, v_j) \in \leq\colon s_i \leq s_j\}$$

und

$$C_V = \{s \in [0,1] \mid \forall (v_i, v_j) \in \leq\colon s_i \leq s_j\}.$$

Mit dieser Notation gilt dann

$$\begin{aligned} \hat{P}_{n_x}^X(A) =& \langle w^x, s^A \rangle \\ \hat{P}_{n_y}^Y(A) =& \langle w^y, s^A \rangle. \end{aligned}$$

Wir betrachten nun die Charakterisierung des Vorliegens von „empirischer stochastischer Dominanz innerhalb der beobachteten Stichproben“ und definieren dazu

$$\begin{aligned} X \leq_{\hat{SD}} Y :& \iff \forall A \in \mathfrak{O}((V, \leq)) : \hat{P}_{n_x}^X(A) \leq \hat{P}_{n_y}^Y(A) \\ &\iff \forall A \in \mathfrak{O}((V, \leq)) : \langle w^x, s^A \rangle \leq \langle w^y, s^A \rangle \\ &\iff \forall A \in \mathfrak{O}((V, \leq)) : \langle w^x, s^A \rangle - \langle w^y, s^A \rangle \leq 0 \\ &\iff \forall A \in \mathfrak{O}((V, \leq)) : \langle w^x - w^y, s^A \rangle \leq 0 \\ &\iff \sup_{A \in \mathfrak{O}((V, \leq))} \langle w^x - w^y, s^A \rangle \leq 0 \\ &\iff \max_{s \in B_V} \langle w^x - w^y, s^A \rangle \leq 0 \\ &\iff^{14} \max_{s \in C_V} \langle w^x - w^y, s^A \rangle \leq 0. \end{aligned}$$

Dies bedeutet: Das Feststellen empirischer stochastischer Dominanz ist lösbar als gewöhnliches lineares Optimierungsproblem über der konvexen Menge C_V. Dafür gibt es Standardmethoden wie das Simplexverfahren oder Innere-Punkte-Verfahren. Das Simplex-Verfahren hat eine exponentielle worst-case Laufzeitkomplexität, jedoch beispielsweise eine polynomiale geglättete Laufzeitkomplexität. Innere-Punkte-Verfahren haben polynomiale Laufzeitkomplexität und sind inzwischen längst auch praktisch mit dem Simplexvarfahren bzw. Varianten davon konkurrenzfähig. Das Optimierungsproblem ist also aus praktischer Sicht auch für größere Probleme lösbar.

Bemerkung 3.5. Jeder Mengenring $\mathcal{MR} \subseteq 2^G$ mit G endlich kann durch (triviale bzw.) einfache Implikationen beschrieben werden. Für eine Optimierung

$$\max_{A \in \mathcal{MR}} \langle w^x - w^y, s^A \rangle$$

kann man daher analog vorgehen. Mit

$$\begin{aligned} B_{\mathcal{MR}} :=& \{s \in \{0,1\}^k \mid \text{ Für jede einfache Implikation } \{g_i\} \to \{g_j\}, \text{ die alle Mengen von} \\ & \mathcal{MR} \text{ respektieren, gilt } s_i \leq s_j \text{ und für jede triviale Implikation } \emptyset \to \{g_i\}, \text{ die} \\ & \text{alle Mengen von } \mathcal{MR} \text{ respektieren, gilt } s_i = 1\} \end{aligned}$$

[14] Dies ist der interessante Schritt, vergleiche Übung.

bzw.

$$C_{\mathcal{MR}} := \{s \in [0,1]^k \mid \text{ – „ –}\}$$

gilt:

$$\max_{A \in \mathcal{MR}} \langle w^x - w^y, s^A \rangle = \max_{s \in B_{\mathcal{MR}}} \langle w^x - w^y, s \rangle = \max_{s \in C_{\mathcal{MR}}} \langle w^x - w^y, s \rangle.$$

Die letzte Gleichheit liegt darin begründet, dass die formale Implikationsrelation „$\rightarrow$“ eine Quasiordnung ist, die von einer Quasiordnung zwischen einelementigen Prämissen und Konklusionen erzeugt wird, die (trivialen und die) einelementigen Implikationen reichen also für die Beschreibung des Mengenrings $\mathcal{MR}$ aus und alle Überlegungen der Übungsaufgabe zu den Oberhalbmengen können entsprechend übertragen werden.

Was, wenn man keinen Mengenring hat, sondern nur ein Hüllensystem? Wir hatten beispielhaft angschaut:

$$\begin{aligned} G &= \mathbb{R}^2 \cap \{p_1, \ldots, p_k\} \text{ mit } p_1, \ldots, p_k \text{ fixierte Punkte aus } \mathbb{R}^2. \\ M &\ldots \text{ Menge aller Halbräume von } \mathbb{R}^2. \\ gIm &\text{ falls Punkt } g \text{ in Halbraum } m \text{ liegt.} \end{aligned}$$

Dann waren die Begriffsumfänge die konvexen Mengen von $\mathbb{R}^2$, projiziert auf $\{p_1, \ldots, p_k\}$. Das Hüllensystem aller Begriffsumfänge ist in diesem Beispiel im Allgemeinen nicht abgeschlossen unter Vereinigungen, also wirklich nur ein Hüllensystem und kein Mengenring. Allgemein betrachte einen beliebigen formalen Kontext $\mathbb{K} = (G, M, I)$ und $\mathcal{I}(\mathbb{K})$:= Menge aller Gegenstandsimplikationen, die im Kontext gelten und $J \subseteq \mathcal{I}(\mathbb{K})$ eine Basis von $\mathcal{I}(\mathbb{K})$. Dann kann man das Problem

$$\langle w^x - w^y, s^A \rangle \longrightarrow \max$$

unter $A \subseteq G$ und A respektiert alle Implikationen aus $\mathcal{I}(\mathbb{K})$ bzw. J als **binäres** Optimierungsproblem lösen:

$$\begin{aligned} \langle w^x - w^y, s \rangle &\longrightarrow \max \text{ unter} \\ s &\in \{0,1\}^k \text{ und} \\ \forall (Y, Z) \in J : \sum_{i: y_i \in Y} s_i - \frac{1}{|Z|} \sum_{i: v_i \in Z} s_i &\leq |Y| - 1 \quad (^*). \end{aligned}$$

Die Ungleichungen $(^*)$ modellieren hier genau, dass alle formalen Implikationen respektiert werden. Hier ist wichtig zu bemerken, dass die Forderung $s \in \{0,1\}^k$ im Allgemeinen **nicht** zur Forderung $s \in [0,1]^k$ relaxiert werden kann, d.h., im Allgemeinen muss hier ein NP-schweres Problem gelöst werden, was computational üblicherweise sehr viel aufwendiger ist.

4 (Lineare) Optimierung auf (durch einen formalen Kontext gegebenen) Hüllensystemen: Subgroup Discovery

Definition 4.1 (Subgroup Discovery, [Wrobel, 2001])
"In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer,...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting" i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest." [Wrobel, 2001]

Definition 4.2 (Problemformulierung der Subgroup-Discovery in der Sprache der formalen Begriffsanalyse)
Gegeben sei ein formaler Kontext $\mathbb{K} = (G, M, I)$, eine Zahl $k \in \mathbb{N}$, eine Zielvariable $t : G \longrightarrow \mathcal{A}$ mit Ausprägungen in der Wertemenge $\mathcal{A}$, sowie eine sogenannte **Qualitätsfunktion** $q : 2^G \longrightarrow \mathbb{R}$, die die "Interessantheit" von Subgruppen bezüglich der Zielvariable t quantifiziert. Die Aufgabe der Subgroup-Discovery besteht darin, diejenigen k formalen Begriffe (A, B) zu finden, für die der Wert

$$q_t(A)$$

der Qualitätsfunktion am größten ist.

Bemerkung 4.1. Im Folgenden werden wir uns auf den Fall $k = 1$ und eine binäre Zielvariable t beschränken.

Beipiele für Qualitätsfunktionen

- Qualitätsfunktion von Piatetsky-Shapiro: Für $t : G \longrightarrow \{0, 1\}$ definiere

  $$q_{ps}(A) = n(p - p_0)$$

  mit

  $$\begin{aligned} n &= |A| \\ p_0 &= \frac{|\{g \in G \mid t(g) = 1\}|}{|G|} \\ p &= \frac{|\{g \in A \mid t(g) = 1\}|}{|A|}. \end{aligned}$$

  Dann gilt: $q_{ps} = \langle w^x - w^y, s^A \rangle \cdot C$ mit der Konstanten $C = \frac{|\{g \in G | t(g) = 0\}|}{|G|}$. (Hier ist w^x der Gewichtsvektor für die Population von Einheiten mit Ausprägung der Zielvariablen 1 und w^y ist der Gewichtsvektor der Einheiten mit Ausprägung der Zielvariablen 0.) Also ist q_{ps} eine lineare Funktion im charakteristischen Vektor s^A.

- $q(A) = \begin{cases} \frac{p}{p_0} & \text{falls } n \geq n^* \text{ für ein fixes } n^* \in \mathbb{N} \\ 0 & \text{sonst} \end{cases}$
- $q(A) = n^\alpha \cdot (p - p_0)$ mit $\alpha \in [0, 1]$
- $q(A) = \frac{(p-p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{\frac{N}{N-n}}$ mit $N = |G|$
- $q(A) = \frac{n}{N-n}(p - p_0)^2$

Im Folgenden betrachten wir nur q_{ps}, d.h., wir haben es mit linearer Optimierung auf Hüllensystemen zu tun.

Bemerkung 4.2. Für binäre Zielvariablen wurden in der Literatur die folgenden Struktureigenschaften einer Qualitätsfunktion als wünschenswert herausgestellt:

Monotoniaxiome für Qualitätsfunktionen bei binären Zielvariablen (cf.,[Piatetsky-Shapiro, 1991, Major and Mangano, 1995]):

1. $q(A) = 0$ für $p = p_0$,
2. q ist isoton in p für fixiertes n.
3. q ist antiton in in n für $p = c/n$ mit fixem c,
4. q ist isoton in n, falls $p > p_0$ fixiert wird.

Frage: Wie kann man effizient die „interessanteste“ Subgruppe berechnen?

- Wenn $\mathfrak{B}((G, M, I))$ klein ist, dann exhaustives Durchsuchen des gesamten Suchraumes beispielsweise nach vorherigem explizitem Berechnen aller formalen Begriffe (A, B) sowie des Wertes der Qualitätsfunktion $q(A)$.
- In „klassischer“ Subgroup Discovery werden formale Begriffe durch sogenannte subgroup descriptions über die Ausprägung von Merkmalen beschrieben. Z.B. für kategoriale Kovariablen $X_1, \ldots, X_n$ mit Ausprägungen z.B. aus $\{a, b, c\}$ ist eine subgroup description dann eine Beschreibung einer Subgruppe über bestimmte Ausprägungsspezifikationen, z.B. „Menge aller Gegenstände/statistischen Einheiten mit Ausprägung $x_1 = a$ & $x_3 = b$ & ...“. Eine naive exhaustive Suche würde einfach alle subgroup descriptions explizit betrachten.
- Achtung: verschiedene subgroup descriptions können zu den gleichen Gegenstandsmengen führen $\rightsquigarrow$ Viel unnötige Redundanz $\rightsquigarrow$ Viele computationale Tricks, um diese Redundanz zu reduzieren:
- Eine Methode unter vielen anderen: Formale Begriffsanalyse, Anwendung beispielsweise des next-closure Algorithmus' [Ganter, 2013, S.84-89] (Hier werden nicht alle subgroup descriptions betrachtet, sondern es werden direkt alle formalen Begriffe enumeriert, es wird also jeder formale Begriff nur einmal berechnet, d.h., subgroup descriptions, die zum selben Begriff führen, werden wirklich nicht mehrmals berechnet.)
- Wenn es sehr viele bzw. zu viele formale Begriffe gibt: Heuristiken (z.B. Beam search).
- Andere Möglichkeit: Exhaustive Suche mit pruning des Suchraumes, z.B. mit Hilfe von optimistic estimates: Baue Suchraum von größeren Subgruppen beschrieben durch kleinere subgroup descriptions hin zu kleineren Subgruppen beschrieben durch größere subgroup descriptions hin auf. (Kleiner bzw. größer sind hier im Sinne von Mengeninklusion zu verstehen.) In einigen Fällen (z.B. für die Qualitätsfunktion von Piatetsky-Shapiro) ist das Folgende möglich: Für eine Subgruppe mit subgroup description z.B. „$x_1 = a$ & $x_5 = c$“ betrachte alle möglichen spezifischeren Subgruppen, also Subgruppen mit subgroup description „$x_1 = a$ & $x_5 = c$ & ...“. Beobachtung: Alle Ggenstände in der betrachteten

Subgruppe haben Ausprägung 1 oder 0 für die Zielvariable. Das beste, was bei einer Verkleinerung der betrachteten Subgruppe passieren kann, ist, dass alle Gegenstände mit Ausprägung 0 aus der Subgruppe herausfallen (und gleichzeitig alle weiteren Gegenstände mit Ausprägung 1 in der Subgruppe verbleiben). Diese Betrachtung erlaubt eine Abschätzung (sogenannter optimistic estimate) darüber, wie groß die Qualitätsfunktion bei einer Verkleinerung der Subgruppe maximal werden kann. Der so erhaltene optimistic estimate erlaubt ein pruning des Suchraumes: Subgruppen, für die der optimistic estimate kleiner ist als der maximale Wert der Qualitätsfunktion über alle bereits inspizierten Subgruppen, muss man nicht weiter inspizieren.

Jetzt nochmal zurück zur formalen Begriffsanalyse: Eine besonders elegante Art und Weise, das Subgroup Discovery Problem anzugehen, ist ganz im Geiste der formalen Begriffsanalyse und der Peirce'schen Minimalforderung nach **sowohl** der Klarheit, **als auch** der Deutlichkeit der Gedanken gedacht: Wir werden dazu sowohl den Begriffsumfang, als auch den Begriffsinhalt **gleichzeitig** zur Beschreibung eines formalen Begriffes heranziehen. Rekapitulieren wir dazu nochmal die Eigenschaft, formaler Begriff zu sein. Ein Paar (A, B) ist ein formaler Begriff, falls gilt:

i) $\forall g \in A \forall m \in B : gIm$
oder anders ausgedrückt: $\forall (g,m) \in (G \times M) \backslash I : g \notin A$ oder $m \notin B$
oder noch anders: $\forall (g,m) \in (G \times M) \backslash I : \underbrace{g \in A \Longrightarrow m \notin B}_{(\Longleftrightarrow m \in B \Longrightarrow g \notin A)}$

ii) $\forall m \in M : m \notin B \Longrightarrow \exists g \in A : g \not I m$

iii) $\forall g \in G : g \notin A \Longrightarrow \exists m \in B : g \not I m.$

Diese Bedingungen lassen sich leicht in Ungleichungsnebenbedingungen innerhalb eines binären Programms zur Lösung des Subgroup Discovery Problems umwandeln. Betrachte dazu für einen formalen Begriff (A, B) einen zugehörigen charakteristischen Vektor s mit $|G| + |M|$ Komponenten, indiziert mit den Mengen G und M und der Kodierung
$s_g = \begin{cases} 1 & \text{falls } g \in A \\ 0 & \text{sonst} \end{cases}$ und $s_m = \begin{cases} 1 & \text{falls } m \in B \\ 0 & \text{sonst} \end{cases}$. Dann kann man die Bedingungen $i) - iii)$ übersetzen zu:

$\tilde{i})$ $\forall (g,m) \in (G \times M) \backslash I : s_g + s_m \leq 1$

$\tilde{ii})$ $\forall m \in M : \sum_{g \in G : g \not I m} s_g + s_m \geq 1$

$\tilde{iii})$ $\forall g \in G : \sum_{m \in M : g \not I m} s_m + s_g \geq 1.$

Dies sind maximal $|G| \cdot |M| + |G| + |M|$ Ungleichungsnebenbedingungen. Damit kann für die Piatetsky-Shapiro-Qualitätsfunktion das Subgroup Discovery Problem als binäres Optimierungsproblem der Form

$$\begin{aligned} &\langle u, s \rangle \longrightarrow \max \text{ unter den Nebenbedingungen} \\ &\quad \tilde{i}) + \tilde{ii}) + \tilde{iii}) \text{ und} \\ &\quad s \in \{0,1\}^{|G|+|M|} \end{aligned}$$

gelöst werden. Analysiert man die Ganzzahligkeitsbedingungen in obigem binären Optimierungsproblem weiter, so macht man schnell die folgende Beobachtung: Wenn man $s_g \in \{0,1\}$ für alle $g \in G$ fordert, so kann man die Bedingungen $\forall m \in M : s_m \in \{0,1\}$ zu $\forall m \in M : s_m \in [0,1]$ relaxieren. Alternativ kann man auch $\forall m \in M : s_m \in \{0,1\}$ fordern und die Ganzzahligkeitsbedingungen $\forall g \in G : s_g \in \{0,1\}$ zu $\forall g \in G : s_g \in [0,1]$ relaxieren. Grund: Fordert man beispielsweise $\forall m \in M : s_m \in \{0,1\}$ und lediglich $\forall g \in G : s_g \in [0,1]$, so sind für jeden zulässigen Vektor s alle Komponenten s_g automatisch binär: Wäre nämlich für ein zulässiges s und $g \in G$ die Komponente s_g aus $]0,1[$, so müsste, um die Bedingung $\tilde{\text{i}}\tilde{\text{i}}\tilde{\text{i}}$ zu erfüllen, ein $m \in M$ mit $g \not I m$ und $s_m = 1$ existieren. Da aber auch für dieses m die Bedingung $\tilde{\text{i}}$ gelten muss, muss wiederum $s_g = 1$ sein, was ein Widerspruch ist, also kann $s_g \in]0,1[$ für ein zulässiges s nicht vorkommen. Dies bedeutet insgesamt: Bei Betrachtung der Piatetsky-Shapiro-Qualitätsfunktion kann man das Problem der Subgroup Discovery als gemischt-ganzzahliges lineares Optimierungsproblem (mixed integer linear programming, MILP) mit $|M|$ binären Variablen und $|G|$ Variablen aus $[0,1]$ (bzw. $|G|$ binären Variablen und $|M|$ Variablen aus $[0,1]$) und maximal $|M| \cdot |G| + |M| + |G|$ Nebenbedingungen formulieren. Insbesondere wenn entweder $|M|$ oder $|G|$ nicht zu groß ist, kann das Problem auch noch für relativ große Kontexte gelöst werden. Betrachten wir nun ein kleines

Anwendungsbeispiel zur Illustration

- Allbus 2018 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften), Zusatzbefragung „Soziale Netzwerke und soziale Unterstützung“
- 10 Kovariablen: Antwort auf die Frage „Kennen Sie eine Frau oder einen Mann mit folgenden Berufen?“ (Erklärender Text: „Hier sehen Sie eine Liste von Berufen, in denen vielleicht Freunde, Verwandte oder Bekannte von Ihnen arbeiten. Es geht um alle Menschen, die Sie namentlich und gut genug kennen, um mit ihnen Kontakt aufzunehmen. Wenn Sie mehrere Menschen kennen, die in einem dieser Berufe arbeiten, berücksichtigen Sie bitte nur DIE PERSON, DER SIE SICH AM NÄCHSTEN FÜHLEN. Jeder dieser Berufe könnte sowohl von einer Frau als auch von einem Mann ausgeführt werden.“)

  1. Bus-/LKW-Fahrer
  2. Geschäftsführung
  3. Reinigungskraft
  4. Friseur/in
  5. Leiter Personalabteilung
  6. Rechtsanwalt
  7. Automechaniker/in
  8. Krankenpfleger/in
  9. Polizist/in
  10. Lehrer/in

  mit Anwortmöglichkeiten

  A Familienmitglied oder Verwandte/r
  B Guter Freund/gute Freundin
  C Bekannte/r
  D Niemand

- Interessierende Zielvariable: Antwort auf die Frage „Wie häufig schließen Sie bei diesen Gelegenheiten [mit 3 oder mehr Freunden oder Bekannten etwas essen oder trinken gehen] neue Freundschaften? mit Antwortskala

  1. Nie
  2. Selten
  3. Manchmal
  4. Oft
  5. Sehr oft

  Wir betrachten hier die daraus abgleitete binäre Zielvariable
  $$t = \begin{cases} 1 & \text{falls Antwort „oft“ oder „sehr oft“} \\ 0 & \text{sonst} \end{cases}$$

- Insgesamt $|G| = 1354$ Personen. (Personen, die nicht an der Zusatzbefragung „Soziale Netzwerke und soziale Unterstützung“ teilgenommen haben und Personen, die mit „Keine Angabe“ oder „Kann ich nicht sagen“ geantwortet haben, wurden hier von der Analyse ausgeschlossen.) Von den 1354 Personen hatten 70 Personen Ausprägung 1 für die Zielvariable (also $p_0 \approx 0.052$).

Frage: Wie soll man konkret die 10 Kovariablen begrifflich skalieren? Hier gibt es sehr viele als natürlich erscheinende Möglichkeiten. Aus rein statistischer Sicht (im Sinne von statistischer Inferenz, nicht im Sinne von messtheoretischen Überlegungen) sollte man wohl so skalieren, dass das entstehende Hüllensystem von Begriffsumfängen auf der einen Seite nicht zu klein ist, damit es genügend Subgruppen gibt und man eventuelle systematische Unterschiede bezüglich der Verteilung der Zielvariable gut durch Betrachtung geeigneter Subgruppen entdecken kann. Auf der anderen Seite sollte das Hüllensystem nicht zu groß sein, damit eventuell entdeckte Unterschiede auch statistisch signifikant sind im Sinne, dass man statistisch belastbar von den entdeckten Unterschieden in der Subgruppe der Stichprobe auf Unterschiede in der betrachteten Gesamtpopulation schließen kann. Wir betrachten im Folgenden 4 begriffliche Skalierungen, die zu verschieden großen Hüllensystemen führen:

S1: Nominale Skalierung

S2: Kontranominale Skalierung

S3: Nominale Skalierung plus das zusätzliche Merkmal „$\neq$ Niemand“

S4: Interordinale Skalierung mit Anordnung der Antwortmöglichkeiten als $A < B < C < D$

Dann gilt, wie man sich mit ein bisschen Ruhe überlegt, für die zugehörigen Hüllensysteme der Begriffsumfänge $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ und $\mathcal{S}_4$ die Beziehung

$$\mathcal{S}_1 \subseteq \mathcal{S}_3 \subseteq \mathcal{S}_4 \subseteq \mathcal{S}_2.$$

Bevor wir die Ergebnisse für die verschiedenen Skalierungen besprechen, wollen wir kurz analysieren, wie das Hüllensystem aller Begriffsumfänge bei kontranominaler Skalierung aller 10 Variablen konkret aussieht. (Die Kontranominalskala wird später bei der Extremaltheorie für Begriffsverbände bzw. innerhalb der Vapnik-Chervonenkis Theorie im Kontext von Begriffsverbänden eine prominente Rolle spielen. Für die anderen Skalierungen möge man sich ebenfalls vergegenwärtigen, wie das Hüllensystem aller Begriffsumfänge konkret aussieht,

einfach um zu wissen, was man hier tut.) Betrachten wir also beispielsweise die erste Kovariable (Bekanntschaft Bus-/LKW-Fahrer). Dann sieht der Kontext bezüglich dieser Kovariable(, sofern alle möglichen Ausprägungen auch beobachtet wurden,) so aus:

	$\neq A$	$\neq B$	$\neq C$	$\neq D$
g_1	◯	x	x	x
g_2	x	◯	x	x
g_3	x	x	◯	x
g_4	x	x	x	◯
⋮	⋮	⋮	⋮	⋮

Welche Subgruppen werden nun durch eine kontranominale Skalierung erfasst, d.h., wie sehen die Begriffumfänge bzw. -Inhalte aus? Die Begriffsinhalte sind genau alle Schnitte von Begriffsinhalten von Gegenstandsbegriffen. Beobachtung: Für jedes Merkmal m existiert ein Gegenstand g_m mit zugehörigem Gegenstandsbegriff $B_m = (\{g_m\}'', \{g_m\}')$, für den der Begriffsinhalt genau das Merkmal m nicht enthält, ansonsten aber alle anderen Merkmale enthält. Für eine beliebige Merkmalsmenge A ist somit

$$A = \bigcap_{m \notin A} \{g_m\}'.$$

Damit sind alle beliebigen Merkmalsmengen Begriffsinhalte. (Dies ist im Sinne einer lokalen Betrachtung von jeweils einer der 10 Kovariablen gemeint.) Für jede der 10 Kovariablen mit 4 Ausprägungen würde man damit jeweils alle möglichen 2^4 Spezifikationen der Ausprägung der Kovariable (also z.B. $x_1 \in \{A, D\}$) zur Definition einer Subgruppe heranziehen. Formal gesehen hätte der Kontext $10 \cdot 4 = 40$ Spalten und somit maximal $2^{40} \approx 1 \cdot 10^{12}$ formale Begriffe. (Man beachte aber, dass dies nicht alle möglichen Spezifikationen von Kovariablenausprägungen sind, pro Dimension betrachtet man zwar alle möglichen Spezifikationen, also alle Teilmengen $T \subseteq 2^{\{A,B,C,D\}}$, insgesamt betrachtet man aber nur 10-fache kartesische Produkte solcher Teilmengen, also 10-dimensionale „Rechteckmengen“. Zum Vergleich: Eine 10-dimensionale Kovariable mit jeweils 4 möglichen Ausprägungen pro Dimension kann insgesamt maximal $4^{10} = 1048576$ Ausprgungen annehmen, d.h., man könnte insgesamt bis zu $2^{(4^{10})} = 2^{1048576} >> 2^{40}$ Subgruppenspezifikationen betrachten. Natürlich ist wegen $|G| = 1354$ die Anzahl der konkret erhaltenen Subgruppen dann auf $2^{1354} >> 2^{40}$ beschränkt.) Die Ergebnisse der Subgroup Discovery für die verschiedenen begrifflichen Skalierungen sind in der folgenden Tabelle zusammengefasst:

	$S1$: nominal	$S2$: kontranominal	$S3$: nominal + „$\neq$ niemand“	$S4$: interordinal
n	209	234	422	258
p	0.1	0.13	0.09	0.12
$ps = n(p - p_0)$	10.19	18.9	16.18	16.66
$argmax$	Friseur/in=Freundin	Reinigungskraft ≠ Bekannte/r & Friseurin ≠ Bekannter/r & Leiter Personalabt. ≠ Niemand	Leiter Personalabt. ≠ Niemand & Krankenpfleger/in ≠ Niemand & Automechaniker/in ≠ Niemand	Leiter Personalabt. ≤ Bekannte/r Krankenpfleger/in ≤ Freund/in & Automechaniker/in ≤ Bekannte/r
p^*	≈ 0.15	≈ 0.05	≈ 0.002	≈ 0.05
V.C.-Dimension	10	25	10	15

Nun zum Hintergrund der oben angegebenen p-Werte, die in der Tabelle mit p^* bezeichnet wurden. Diese basieren auf einem Permutationstest: Permutiere zufällig die Labels der Zielvariable und berechne den Wert der Qualitätsfunktion ps für den permutierten Datensatz. Die

Verteilung der Werte ps entspricht dann genau der **bedingten** Verteilung von ps bedingt auf die beobachteten Kovariablenausprägungen unter der Nullhypothese, dass die Verteilung der Zielvariable nicht von der Ausprägung der Kovariablen abhängt. Achtung: Ist $p^* \leq \alpha$ für gewähltes Signifikanzniveau α, so ist „die Situation statistisch signifikant von H_0 verschieden", aber: Damit ist noch nichts direkt über die statistische Belastbarkeit der mit der entdeckten Subgruppe assoziierten Aussage über das „Wie" der Abweichung von H_0 gesagt. Beispielsweise ist die Aussage (hättte man a priori Skalierung S_3 gewählt) „In der Subgruppe ... ist der Anteil derer, die oft oder sehr oft neue Bekanntschaften schließen, gegenüber den anderen betrachteten Subgruppen am größten." **nicht** statistisch belastbar. Abschließend sei noch eine sehr verknappte, skizzenhafte Einordnung der Subgroup Discovery (in Verbindung mit der formalen Begriffsanalyse, dem begrifflichen Skalieren und einer statistischen Betrachtung über einen Permutationstest) in das statistische Methodenspektrum erlaubt:

i) Die Subgroup Discovery könnte man als eher nichtparametrische Methode betrachten, für die Validität des Permutationstests wäre lediglich eine *i.i.d.*-Annahme nötig.

ii) Insbesondere gibt es keine Homogenitätsannahmen[15] über das unterliegende Wahrscheinlichkeitsgesetz, wie es beispielsweise bei einer Analyse über einen Modellierungsansatz und eine z.B. (generalisierte) lineare Regression im nicht-saturierten Fall der Fall wäre.

iii) Wesentlich ist, dass man in gewissem Sinne ausgewählte und insbesondere nicht zu viele Ereignisse/Subgruppen gleichzeitig anschaut. Die statistische Rechtfertigung für die Sinnhaftigkeit eines Vorgehens im Sinne der Subgroup Discovery unter Beachtung obiger Selbstbeschränkung liefert die Theorie gleichmäßiger Konvergenz empirischer Wahrscheinlichkeitsmaße zu den unterliegenden unbekannten wahren Wahrscheinlichkeitsgesetzen bezüglich von Klassen $\mathcal{S}$ von Ereignissen, sofern diese Klassen nicht „zu groß" sind, vergleiche den späteren Abschnitt zur statistischen Lerntheorie/Vapnik-Chervonenkis Theorie und zur Regularisierung. Hier sei schonmal die folgende Bemerkung gestattet: Für ein Ereignis A gilt

$$P(|\hat{P}_n(A) - P(A)| \leq \varepsilon) \geq 1 - \delta$$

für beliebige $\delta, \varepsilon > 0$ falls n groß genug ist. (Hier ist P das wahre, unbekannte unterliegende Wahrscheinlichkeitsbildmaß und $\hat{P}_n$ ist das empirische Maß einer i.i.d.-Stichprobe der Größe n, d.h., $\hat{P}_n(A)$ ist einfach der Anteil der beobachteten Werte in der Stichprobe, die in der Menge A liegen.) Wichtig ist hier, dass für beliebiges $\varepsilon > 0$ gilt:

$$P(|\hat{P}_n(A) - P(A)| \leq \varepsilon) \leq Ce^{-Dn}$$

mit Konstanten $C, D > 0$, d.h., die rechte Seite fällt exponentiell schnell in n. Für eine endliche Familie $\mathcal{S}$ von Ereignissen gilt nun:

[15] Damit ist hier keine Homoskedastizitätsannahme gemeint, sondern die Strukturannahme, dass der „Effekt" von Kovariablen nicht von der Ausprägung anderer Kovariablen abhängt, dass der Effekt von Kovariablen also homogen über dem Kovariablenraum ist.

$$\begin{aligned} P(\sup_{A\in\mathcal{S}} |\hat{P}_n(A) - P(A)| \geq \varepsilon) = & \quad P(\exists A \in \mathcal{S} : |\hat{P}_n(A) - P(A)| \geq \varepsilon) \\ = & \quad P\left(\bigcup_{A\in\mathcal{S}} \{|\hat{P}_n(A) - P(A)| \geq \varepsilon\}\right) \\ \overset{\text{union bound}}{\leq} & \quad \sum_{A\in\mathcal{S}} P(\{|\hat{P}_n(A) - P(A)| \geq \varepsilon\}) \\ \leq & \quad |\mathcal{S}| \cdot Ce^{-Dn}, \end{aligned}$$

d.h., für $\mathcal{S}$ endlich konvergiert das emirische Maß $\hat{P}_n$ gleichmäßig auf $\mathcal{S}$ gegen das wahre Maß P. Außerdem würde auch für ein potentiell unendliches Mengensystem $\mathcal{S}$, das in n nur polynomial wächst, ebenfalls

$$\sup_{A\in\mathcal{S}} |\hat{P}_n(A) - P(A)| \xrightarrow{\mathbb{P}} 0$$

gelten.

iv) Das begriffliche Skalieren eröffnet viele Möglichkeiten, das Mengensystem $\mathcal{S}$ sinnvoll und nicht zu groß (bzw. auch nicht zu klein) zu gestalten, vergleiche auch den späteren Abschnitt zur Regularisierung. Aber: Selbstverständlich muss man sich darum kümmern.

v) Insbesondere ist es recht einfach, bestimmte Aspekte explizit mit einzubeziehen, z.B. die Zusammenfassung von Merkmalsausprägungen zu bestimmten Kategorien (z.B. Merkmal $X \in \{b, d\}$).

vi) Technischer Aspekt: Es ist möglich, beinahe beliebige constraints in die Optimierung mit aufzunehmen, z.B.:

- $n \geq n^*$ (oder „=“ oder „≤“), wobei n die Größe der Subpopulation und n^* eine feste natürliche Zahl ist.
- Oder betrachte nur Begriffe (A, B), die $|B| \leq C$ (oder $|A| \geq C$, das ist genau der Fall oben) mit einer festen Schranke $C \in \mathbb{N}$ erfüllen.
- Oder für eine weitere Zielvariable/Kontrollvariable, die „linear in s ist“, fordere

  $$\langle v, s\rangle \geq C,$$

  wobei hier v ein Vektor ist, der die weitere Zielvariable/Kontrollvariable modelliert und C eine feste Konstante ist.
- Oder fordere, dass zusätzliche Implikationen (Gegenstandsimplikationen, Merkmalsimplikationen oder auch gemischte Implikationen) von den betrachteten formalen Begriffen erfüllt werden.

(Das Fordern zusätzlicher constraints kann auch im Sinne einer Regularisierung des Problems (vergleiche später) verstanden bzw. benutzt werden.)

5 Was Begriffsverbände groß macht: Extramaltheorie für Begriffsverbände

Sei $\mathbb{K} = (G, M, I)$ ein (endlicher) Kontext mit (O.B.d.A.) $|G| \geq |M|$. Wie groß kann $\mathfrak{B}(\mathbb{K})$ dann maximal sein? Natürlich ist $|\mathfrak{B}(\mathbb{K})| \leq 2^{|M|}$, denn es kann maximal $2^{|M|}$ Begriffsinhalte geben. Außerdem kann mit $\mathbb{K} := (2^M, M, \ni)$ der Extremfall $|\mathfrak{B}(\mathbb{K})| = 2^{|M|}$ erreicht werden. Die Frage ist nun, ob es auch kleinere Kontexte (im Sinne von möglicht wenige Gegenstände) $\mathbb{K}$ mit $|\mathfrak{B}(\mathbb{K})| = 2^{|M|}$ gibt. Überlegungen dazu:

- Alle Begriffsinhalte sind Schnitte von Inhalten von Gegenstandsbegriffen.
- Für $|\mathfrak{B}(\mathbb{K})| = 2^{|M|}$ ist nötig, dass alle $|M|-1$-elementigen Merkmalsmengen auch Inhalte sind.
- Verschiedene Gegenstände mit genau gleichen Merkmalen sind redundant in dem Sinne, dass die Entfernung eines von zwei oder mehr Gegenständen mit gleichen Merkmalen das Hüllensystem aller Begriffsinhalte nicht verändert (, wohl aber das System der Begriffsumfänge).
- Redundante Gegenstände und Gegenstände, die alle Merkmale aus M besitzen einmal beiseite gelassen, kann jeder $|M|-1$-elementige Inhalt nur als Inhalt eines Gegenstandsbegriffes erhalten werden. (Wenn man die Inhalte von zwei oder mehr verschiedenen Gegenstandsbegriffen schneidet, dann hat man schon weniger als $|M|-1$ Merkmale im Schnitt.)
- Damit sind für die Erzeugung aller $|M|$ Inhalte mit jeweils $|M|-1$ Elementen mindestens die folgenden Gegenstände nötig:

	m_1	m_2	m_3	$\dots$	$m_{\lvert M\rvert-1}$	$m_{\lvert M\rvert}$
g_1	◯	x	x	$\dots$	x	x
g_2	x	◯	x	$\dots$	x	x
g_3	x	x	◯	$\dots$	x	x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$g_{\lvert M\rvert-1}$	x	x	x	$\dots$	◯	x
$g_{\lvert M\rvert}$	x	x	x	$\dots$	x	◯

- In dieser Situation ist nun bereits jede beliebige Merkmalsmenge $A \subseteq M$ als Schnitt der Inhalte aller derjenigen Gegenstandsbegriffe, deren Inhalte jeweils genau ein Element aus $M \backslash A$ nicht enthalten, darstellbar:

$$A = \bigcap_{m \in M \backslash A} \{g_m\}'.$$

  (Hier ist g_m derjenige Gegenstand, der genau alle Merkmale bis auf das Merkmal m besitzt.)
- Also ist genau für $|G| = |M|$ und $\mathbb{K} = (G, M, I)$ eine Kontranominalskala(, d.h., $\forall g \in G \exists! m_g \in M : g \not I m$ und es gilt $g \neq \tilde{g} \Longrightarrow m_g \neq m_{\tilde{g}}$) ein Kontext mininmaler Größe und $|\mathfrak{B}(\mathbb{K})| = 2^{|M|}$ gegeben.

Wenn wir jetzt einen beliebigen Kontext $\mathbb{K}$ betrachten, könnten wir die Vermutung haben, dass große Subkontexte $\tilde{\mathbb{K}} = (\tilde{G}, \tilde{M}, \tilde{I})$ mit $\tilde{G} \subseteq G, \tilde{M} \subseteq M, \tilde{I} = I \cap G \times M$, die eine Kontranominalskala bilden, einen Begriffsverband groß machen. Sei $\tilde{\mathbb{K}}$ der größte Subkontext, der eine Kontranominalskala ist. Dann haben wir bereits

$$|\mathfrak{B}((G, M, I))| \geq 2^{|\tilde{G}|} = 2^{|\tilde{M}|}.$$

Es könnte jetzt aber sein, dass nicht große Subkontexte, die Kontranominalskalen sind, den Begriffsverband besonders groß machen, sondern dass andere Strukturen, die nicht Kontranominalskalen sind, viel mehr zur Größe des Begriffsverbandes beitragen, man beachte, dass ja der Fall $|\tilde{G}| << |G|$ und $|\tilde{M}| << |M|$ vorkommen kann. Erstaunlicherweise kann man mit der einfachen Charakteristik $K := |\tilde{G}|$ die Größe eines Begriffsverbandes abschätzen zu:

$$|\mathfrak{B}((G, M, I))| \leq \sum_{i=0}^{K} \binom{|M|}{i} \quad (\leq |M|^K + 1) \qquad (^*)$$

$$|\mathfrak{B}((G, M, I))| \leq \sum_{i=0}^{K} \binom{|G|}{i} \quad (\leq |G|^K + 1) \qquad (^{**})$$

(Obige Ungleichungen gelten auch für den Fall $|G| \not\geq |M|$.) Bevor wir zu den obigen Ungleichungen kommen (Abschnitt 6), zunächst noch zur „statistischen Inferenz“ bezogen auf die vorher betrachtete Supremumstatistik T (bzw. die Qualitätsfunktion von Piatetsky-Shapiro im Kontext der Subgroup Discovery): Existiert ein Subkontext $\tilde{\mathbb{K}} = (\tilde{G}, \tilde{M}, \tilde{I})$ der Größe $K = |\tilde{G}| = |\tilde{M}|$, der Kontranominalskala ist, und zieht man aus G eine Stichprobe der Größe K, dann ist das schlimmste, was passieren kann, dass man genau die Gegenstände aus $\tilde{G}$ zieht. Dann wäre nämlich unabhängig davon, welcher Gegenstand zu welcher Subpopulation gehört (bzw. unabhängig von der Zielvariable t im Kontext der Subgroup Discovery) die Teststatistik gleich 1 (bzw. die Qualitätsfunktion von Piatetsky-Shapiro maximal).

Kommen wir nun endlich zu einem bisschen „richtiger Statistik“, wir werden in ganz groben Zügen ein paar Grundlagen der sogenannten Vapnik-Chervonenkis Theorie (bzw. der statistischen Lerntheorie) kennenlernen. In der Vapnik-Chervonenkis Theorie, die wesentlich von Vladimir Vapnik und Alexey Chervonenkis in den 1970-ern entwickelt wurde, geht es darum, unter welchen Voraussetzung ein unterliegendes Wahrscheinlichkeitsgesetz in einem gleichmäßigen Sinne über eine Klasse $\mathcal{S}$ von Ereignissen gut durch das empirische Maß geschätzt werden kann.

6 Elemente der statistischen Lerntheorie/Vapnik-Chervonenkis Theorie

"According to Kolmogorov, in the space of problems suggested by the real world there is a huge subspace where one can find trivial solutions. There is also a huge subspace where solutions are inaccessible. Between these two subspaces there is a tiny subspace where one can find non-trivial solutions. Mathematics operates inside this subspace. It is therefore a big achievement when one can suggest a problem setting and a resolution to this setting and also invent concepts and rules that make proofs both nontrivial and accessible (this is interesting for mathematicians). In order to transform a problem from an inaccessible one to one that has a mathematical solution very often one must simplify the setting of the problem, perform mathematical analysis, and then apply the result of this analysis to the nonsimplified real-life problem." [Vapnik, 2006]

Definition 6.1 (Projektion, growth-function, shatterable set)
Sei $\mathcal{S} \subseteq 2^V$ ein Mengensystem auf der Grundmenge V. Definiere für eine **beliebige** Menge $A \subseteq V$ (A muss nicht aus $\mathcal{S}$ sein) die Projektion von $\mathcal{S}$ auf A als:

$$\mathcal{S}_A := \{A \cap B \mid B \in \mathcal{S}\}.$$

Die sogenannte growth-function (Wachstumsfunktion) $m^{\mathcal{S}}$ ist definiert als

$$m^{\mathcal{S}} : \mathbb{N} \longrightarrow \mathbb{N} : r \mapsto \max_{\substack{A \subseteq V, \\ |A|=r}} |\mathcal{S}_A|.$$

Für eine **beliebige** Menge $A \subseteq V$ sagen wir, dass A shatterable (zerschmetterbar) bezüglich $\mathcal{S}$ ist, wenn

$$\mathcal{S}_A = 2^A$$

gilt, d.h., wenn alle Teilmengen von A als Schnitte von A mit Mengen aus $\mathcal{S}$ erzeugbar sind. Mit $sh_V(\mathcal{S})$ bezeichnen wir die Menge aller bezüglich $\mathcal{S}$ zerschmetterbaren Teilmengen von V.

Bemerkung 6.1. Es gilt:

$$\mathcal{S}_A \subseteq 2^A;\ |\mathcal{S}_A| \leq 2^{|A|}$$
$$m^{\mathcal{S}}(r) \leq 2^r.$$

Hintergrund der Betrachtung nicht des gesamten Mengensystems $\mathcal{S}$, sondern des projizierten Mengensystems $\mathcal{S}_A$ ist die Überlegung, dass es in gewissem Sinne ausreicht, nur die Elemente des unterliegenden Grundraumes V zu betrachten, die tatsächlich beobachtet wurden. Die Menge A wird also in Gedanken immer die Menge aller in der Stichprobe beobachteten Ausprägungen sein. Ziel der nun folgenden Überlegungen ist es, die Mächtigkeit eines Mengensystems $\mathcal{S}$ mit rein kombinatorischen Mitteln nach oben abzuschätzen. Man kann sich leicht überlegen, dass große zerschmetterbare Mengen das Mengensystem $\mathcal{S}$ groß machen: Ist A shatterable, dann kann jede Teilmenge $B \subseteq A$ als Schnitt von A und einer Menge aus $\mathcal{S}$ dargestellt werden, woraus bereits $|\mathcal{S}| \geq |2^A|$ folgt. Wir sind aber an einer Art Umkehrung dieser Aussage interessiert und wollen nun der folgenden Spekulation nachgehen:

Vermutung. Wenn ein Mengensystem $\mathcal{S}$ (auf einem endlichen Grundraum) sehr groß ist, dann muss es eine große zerschmetterbare Menge (bezüglich $\mathcal{S}$) geben, bzw. umgekehrt: Gibt es

keine große zerschmetterbare Menge (bezüglich $\mathcal{S}$), so ist $\mathcal{S}$ nicht zu groß.

Wenn dem so wäre, dann könnten wir unter Umständen die growth-function $m^{\mathcal{S}}$, die die maximale Mächtigkeit des auf eine endliche Menge der Kradinaliät r projizierte Mengensystems $\mathcal{S}$ beschreibt, geeignet abschätzen. Dies wird sowohl zur Abschätzung der Größe eines Begriffsverbandes, wie auch zu einer statistischen Analyse des Konvergenzverhaltens von den oben betrachteten Supremumsstatistiken dienlich sein. Erstaunlicherweise gilt in der Tat eine sehr verblüffende kombinatorische Relation zwischen der Mächtigkeit eines endlichen Mengensystems $\mathcal{S}$ und der Mächtigkeit der Menge der zerschmetterbaren Mengen von $\mathcal{S}$, die wir zur Beschränkung der growth-function nutzen können:

Satz 6.2 (Mächtigkeit eines Mengensystems, [Pajor, 1985])
Für eine endliche Menge $V = \{1, \ldots, m\}$ und ein beliebiges Mengensystem $\mathcal{S} \subseteq 2^V$ gilt

$$|\mathcal{S}| \leq |sh_V(\mathcal{S})|. \qquad (*)$$

Beweis. (Nur der Vollständigkeit halber.)
Der Beweis läuft über vollständige Induktion über die Mächtigkeit von V.

Induktionsanfang: Für $m = 1$ gilt (*), denn die leere Menge ist immer zerschmetterbar und für den Fall $\mathcal{S} = \{\emptyset, \{1\}\}$ ist $sh_V(\mathcal{S}) = \{\emptyset, \{1\}\}$.
Induktionsvoraussetzung: Gelte (*) für $\tilde{V} = \{1, \ldots, m-1\}$ und $\tilde{S} = 2^{\tilde{V}}$ beliebig.

Induktionsschritt: Betrachte $V = \{1, \ldots, m\}$ und $S \subseteq 2^V$ sowie

$$\begin{aligned} \mathcal{S}_0 &= \{A \mid A \in \mathcal{S}, m \notin A\} && \subseteq 2^{\tilde{V}} \\ \mathcal{S}_1 &= \{A \backslash \{m\} \mid A \in \mathcal{S}, m \in A\} && \subseteq 2^{\tilde{V}}. \end{aligned}$$

Dann ist $|\mathcal{S}| = |\mathcal{S}_0| + |\mathcal{S}_1|$. Nach Induktionsvoraussetzung ist $|sh_{\tilde{V}}(\mathcal{S}_0)| \geq |\mathcal{S}_0|$ und $sh_{\tilde{V}}(\mathcal{S}_1)| \geq |\mathcal{S}_1|$, woraus

$$|sh_{\tilde{V}}(\mathcal{S}_0)| + |sh_{\tilde{V}}(\mathcal{S}_1)| \geq |\mathcal{S}_0| + |\mathcal{S}_1| = |\mathcal{S}|$$

folgt. Weiterhin ist

$$sh_{\tilde{V}}(\mathcal{S}_0) \cup sh_{\tilde{V}}(\mathcal{S}_1) \subseteq sh_V(\mathcal{S}).$$

Betrachte jetzt ein beliebiges A aus dem Schnitt $sh_{\tilde{V}}(\mathcal{S}_0) \cap sh_{\tilde{V}}(\mathcal{S}_1)$.
Dann ist $m \notin A$ und $A \cup \{m\}$ ist in $sh_V(\mathcal{S})$. Das bedeutet, dass für jedes A, das im Schnitt $sh_{\tilde{V}}(\mathcal{S}_0) \cap sh_{\tilde{V}}(\mathcal{S}_1)$ liegt, die Menge $A \cup \{m\}$ eine weitere zerschmetterbare Menge ist.
Damit folgt aus

$$sh_V(\mathcal{S}) \supseteq sh_{\tilde{V}}(\mathcal{S}_0) \cup sh_{\tilde{V}}(\mathcal{S}_1)$$

das

$$|sh_V\mathcal{S}| \geq |sh_{\tilde{V}}(\mathcal{S}_0)| + |sh_{\tilde{V}}(\mathcal{S}_1)| \geq |\mathcal{S}|$$

gilt.

Obiger Satz impliziert nun das folgende Resultat, das von mehreren Autoren innerhalb verschiedener mathematischer Gebiete (Kombinatorik, Modelltheorie, Statistik) publiziert wurde.

(Die erste Publikation dazu scheint [Vapnik and Chervonenkis, 1968] zu sein, man beachte, dass der Satz von Pajor späteren Datums ist.)

Satz 6.3 ([Vapnik and Chervonenkis, 1968, Sauer, 1972, Shelah, 1972])
Sei $\mathcal{S} \subseteq 2^{\{1,\ldots,m\}}$ derart, dass es keine zerschmetterbare Menge der Kardinalität $k+1$ gibt. Dann ist

$$|\mathcal{S}| \leq \sum_{i=0}^{k} \binom{m}{i} \leq m^k + 1.$$

Beweis (Nur erste Ungleichung.) Da mit jeder zerschmetterbaren Menge A auch jede Teilmenge $B \subseteq A$ zerschmetterbar ist, kann es unter obiger Voraussetzung keine zerschmetterbaren Mengen der Kardinalität größer als k geben. Damit ist

$$sh_V \subseteq \{A \subseteq \{1,\ldots,m\} \mid |A| = i, i \in \{0,\ldots,k\}\}$$

und es folgt

$$|\mathcal{S}| \leq |sh_V(\mathcal{S})| \leq \sum_{i=0}^{k} \binom{m}{i}.$$

Obiges Resultat gibt Anlass zu folgender Definition:

Definition 6.4 (Vapnik-Chervonenkis-Dimension)
Für ein Mengensystem $\mathcal{S} \subseteq 2^V$ mit V nicht notwendigerweise endlich ist die Vapnik-Chervonenkis-Dimension (kurz V.C.-Dimension, in Zeichen: $VC(\mathcal{S})$) von $\mathcal{S}$ definiert als die größte mögliche Kardinalität einer zerschemtterbaren Teilmenge von V. Gibt es zu jedem $r \in \mathbb{N}$ eine zerschmetterbare Menge der Kardinaliät r, so ist die V.C.-Dimension von $\mathcal{S}$ definiert als ∞.

Im Zusammenhang mit der formalen Begriffsanalyse, wo wir Hülllensysteme von Begriffsumfängen bzw. Begriffsinhalten betrachten, ergibt sich außerdem sofort die folgende

Definition 6.5 (V.C.-Dimension eines formalen Kontextes)
Gegeben sei ein formaler Kontext $\mathbb{K} = (G, M, I)$. Da nach Übung 4, Aufgabe 3 die zerschmetterbaren Mengen genau durch die Kontranominalskalen gegeben sind, entspricht die V.C.-Dimension des Hüllensystems aller Begriffsufänge der maximalen Größe $K := |\tilde{G}| = |\tilde{M}|$ eines Subkontextes $\tilde{K} = (\tilde{G}, \tilde{M}, \tilde{I})$, der Kontranominalskala ist. Insbesondere ist damit die V.C.-Dimension des Hüllensystems aller Umfänge gleich der V.C.-Dimension des Hüllensystems aller Inhalte. Diese nennen wir deshalb auch die V.C.-Dimension des Kontextes $\mathbb{K}$, in Zeichen: $VC(\mathbb{K})$.

Satz 6.6 (Hinreichende Bedingung für gleichmäßige Konvergenz)
Definiere die Statistiken

$$D_n := \sup_{A \in \mathcal{S}} |\hat{P}_n^X(A) - P(A)|$$
$$\tilde{D}_n := \sup_{A \in \mathcal{S}} |\hat{P}_n^X(A) - \hat{P}_n^Y(A)|.$$

Ist die V.C.-Dimension von $\mathcal{S}$ endlich, so gilt

$$D_n \overset{P-f.s.}{\longrightarrow} 0$$
$$\tilde{D}_n \overset{P-f.s.}{\longrightarrow} 0.$$

Konkret gilt für $n \geq \frac{2}{\varepsilon^2}$

$$P(D_n \geq \varepsilon) \leq 6 \cdot m^{\mathcal{S}}(2n) e^{-\frac{\varepsilon^2 n}{4}} \leq 6 \cdot ((2n)^{VC(\mathcal{S})} + 1) e^{-\frac{\varepsilon^2 n}{4}}$$
$$P(\tilde{D}_n \geq \varepsilon) \leq 3 \cdot m^{\mathcal{S}}(2n) e^{-\varepsilon^2 n} \leq 3 \cdot ((2n)^{VC(\mathcal{S})} + 1) e^{-\varepsilon^2 n}.$$

Zum Beweis. Betrachte die zufällige Menge

$$B(\omega) := \{X_1(\omega), \dots X_n(\omega), Y_1(\omega), \dots, Y_n(\omega)\}.$$

und das bedingte Wahrscheinlichkeitsmaß $Q := P(\cdot \mid B = b)$ mit b der Menge aller wirklich beobachteten Ausprägungen in der Stichprobe $x_1, \dots, x_n, y_1, \dots, y_n$. Dann ist für beliebiges $\varepsilon > 0$

$$\begin{aligned}
Q(\tilde{D}_n > \varepsilon) &= Q(\sup_{A \in \mathcal{S}} |\hat{P}_n^X(A) - \hat{P}_n^Y(A)| > \varepsilon) \\
&= Q(\sup_{A \in \mathcal{S}_B} |\hat{P}_n^X(A) - \hat{P}_n^Y(A)| > \varepsilon) \\
&= Q(\exists A \in \mathcal{S}_B : |\hat{P}_n^X(A) - \hat{P}_n^Y(A)| > \varepsilon) \\
&= Q(\bigcup_{A \in \mathcal{S}_B} \{|\hat{P}_n^X(A) - \hat{P}_n^Y(A)| > \varepsilon\}) \\
&\overset{\text{union bound}}{\leq} \sum_{A \in \mathcal{S}_B} Q(\{|\hat{P}_n^X(A) - \hat{P}_n^Y(A)| > \varepsilon\}) \\
&\overset{\text{Chernoff Ungleichung}}{\leq} |\mathcal{S}_B| \cdot C \cdot e^{-Dn} \\
&\leq m^S(2n) \cdot C \cdot e^{-Dn} \\
&\leq (2n^{VC(\mathcal{S})} + 1) \cdot C \cdot e^{-Dn} \longrightarrow 0 \text{ für } n \to \infty
\end{aligned}$$

(Hier sind C, D feste positive Konstanten.) Damit folgt auch $P(\tilde{D}_n > \varepsilon) \longrightarrow 0$ für $n \to \infty$. Zu D_n: Man kann zeigen, dass $P(D_n \geq \varepsilon) \leq 2P(\tilde{D}_n \geq \frac{\varepsilon}{2})$ für n hinreichend groß.

Über die obige hinreichende Bedingung hinaus ist es für gleichmäßige Konvergenz auch in folgendem Sinne notwendig, dass die „erwartete V.C.-Dimension“ nicht zu schnell mit dem Stichprobenumfang wächst:

Satz 6.7 (Notwendige (und hinreichende) Konvergenzbedingung)
Eine notwendige (und hinreichende) Bedingung dafür, dass

$$D_n \overset{P-f.s.}{\longrightarrow} 0$$

bzw.

$$\tilde{D}_n \overset{P-f.s.}{\longrightarrow} 0$$

gilt, ist die Bedingung

$$\lim_{n\to\infty} \frac{\mathbb{E}_P(\log_2 |\mathcal{S}_{\{X_1,\ldots,X_n\}}|)}{n} = 0.$$

Während die hinreichende Bedingung verteilungsfrei ist in dem Sinne, dass das Kriterium einer endlichen V.C.-Dimension nicht vom unterliegenden Wahrscheinlichkeitsmaß P abhängt, ist die notwendige Bedingung sehr wohl vom Gesetz P abhängig, es kann also sein, dass man trotz unendlicher V.C.-Dimension gleichmäßige Konvergenz hat. In diesem Sinne wird die V.C. Theorie manchmal als worst-case Theorie bezeichnet (bzw. gelegentlich auch belächelt). Vladimir Vapnik äußerte sich unlängst in einem Interiew dazu folgendermaßen:

"There is a mathematical setting. When I came to [the] United State[s] in 1990 first, people did not know V.C.-theory, they did not know statistical learning theory. In Russia, it was published two monographs, our monographs, but in Amerika they did not know. Then, they learnded it and somebody told me that it is worst-case theory and they will create real-case theory, but till now, they did not. Because it is [a] mathematical tool, you can do only what you can do using mathematics, and which has clear understanding and clear description. And for this reason, we introduced complexity. And we need this, because using .... V.C.-dimension you can prove some theorems...." [Vapnik 2018]

7 Regularisierung

Regularisierung meint meist eigentlich eine mathematische Methode zur Behandlung schlecht gestellter bzw. schlecht konditionierter Probleme. In unserem Zusammenhang soll mit Regularisierung einfach nur die Verkleinerung eines Mengensystems $\mathcal{S}$ zu einem kleineren Mengensystem $\mathcal{T}$ und die anschließende Betrachtung einer Supremumstatistik auf diesem kleineren System zur „Zügelung“ des statistischen Verhaltens der Statistik gemeint sein.

7.1 Regularisierung im Zusammenhang mit Stochastischer Dominanz/Oberhalbmengen

Gegeben sei eine geordnete Menge $(V, \leq)$ und $\mathcal{S} := \mathfrak{O}((V, \leq))$ das Mengensystem aller Oberhalbmengen von $(V, \leq)$. Wenn jetzt $\mathcal{S}$ aus statistischer Sicht zu groß ist, stellt sich die Frage, wie man $\mathcal{S}$ kleiner machen kann. Da $\mathcal{S}$ ja nur implizit gegeben ist (nämlich durch alle formalen Implikationen, die $\mathcal{S}$ respektiert), kann man jedenfalls aus technischer Sicht nicht einfach (willkürlich) ein paar Mengen aus $\mathcal{S}$ entfernen. Für eine endliche geordnete Menge war eine Oberhalbmenge A jedoch beschreibbar über ihre minimalen Elemente als

$$A = \uparrow B$$

mit B der Menge aller minimalen Elemente von A. Für eine Regularisierung könnte man jetzt beispielsweise die folgenden Ideen haben:

a) Betrachte nur Oberhalbmengen, die von Mengen B mit $|B| \leq K$ erzeugt werden. Das Problem dabei ist jedoch, dass das entstehende Mengensystem im Allgemeinen kein Hüllensystem mehr ist und deshalb nicht mehr mit formalen Implikationen beschrieben werden kann. (Man könnte aber trotzdem eine MILP-Formulierung des entstehenden Optimierungsproblems

aufschreiben, was wir hier aber nicht weiter verfolgen wollen.)

b) Betrachte eine Teilmenge $W \subseteq V$ und das zugehörige Mengensystem $\mathcal{T} \subseteq \mathcal{S}$ definiert durch

$$\mathcal{T} = \{\uparrow B \mid B \subseteq W\}.$$

Dann ist $\mathcal{T}$ im Allgemeinen auch kein Hüllensystem. Aber, wenn B einen vollständigen Verband bildet(, was man leicht durch Vervollständigung erreichen kann), dann schon. Man beachte hier, dass Elemente $x \in V \backslash W$ nicht vollständig von der Analyse ausgeschlossen werden, sondern dass sie nur für die Generierung des Hüllensystems, auf dem die Analyse zum Schluss stattfindet, nicht herangezogen werden, wir betrachten nämlich das Hüllensystem

$$\mathcal{T} = \{\{x \in V \mid \exists b \in B : b \leq x\} \mid B \subseteq W\}.$$

Dieses Mengensystem ist sogar ein Mengenring. Allerdings kann $\mathcal{T}$ immer noch zu groß für eine explizite Beschreibung sein. Wir können jedoch, Kraft abstrakter Charakterisierung, $\mathcal{T}$ über alle (trivialen und) einfachen Implikationen, die in $\mathcal{T}$ gelten, beschreiben. Bleibt noch die Frage zu klären, ob es für gegebene Elemente $a, b \in V$ einfach ist zu entscheiden, ob die Implikation $\{a\} \longrightarrow \{b\}$ gilt oder nicht. (Anschließend müsste man nur alle möglichen Paare $a, b \in V$ anschauen, was nicht zu viel ist.) Die Entscheidung, ob $\{a\} \longrightarrow \{b\}$ in $\mathcal{T}$ gilt, ist in der Tat einfach genug zu treffen, insbesondere muss man nicht explizit alle Mengen $A \in \mathcal{T}$ anschauen. Es gilt:

i) Existiert ein $w \in W$ mit $a \in\uparrow w$ & $b \notin\uparrow w$, so gilt $\{a\} \longrightarrow \{b\}$ offensichtlich nicht.

ii) Gilt für alle $w \in W$ die Implikation

$$a \in\uparrow w \Longrightarrow b \in\uparrow w,$$

so wird die Implikation $\{a\} \longrightarrow \{b\}$ in der Tat von **allen** Mengen aus $\mathcal{T}$ respektiert, denn für beliebiges $B \subseteq W$ haben wir

$$a \in\uparrow B \Longrightarrow \exists w \in B : a \geq w \Longrightarrow a \in\uparrow w \Longrightarrow b \in\uparrow w \Longrightarrow b \in\uparrow B.$$

Damit ist das Regularisierungsproblem auf der rein technischen Seite gelöst, allerdings wäre natürlich noch die schwierigere Frage zu beantworten, wie man die Teilmenge $W \subseteq V$ denn nun sinnvoll wählt? (geleitet durch substanzwissenschaftliche Überlegungen, irgendwie datenbasiert, ...?)

7.2 Regularisierung von Begriffsverbänden, die durch einen formalen Kontext gegeben sind

Wir haben oben bzw. in der Übung (Übung 4, Aufgabe 3) gesehen, dass die zerschmetterbaren Mengen genau durch Kontranominalskalen gegeben sind. Dies legt nahe, zunächst nach großen zerschmetterbaren Mengen zu suchen und diese anschließend in einem geigneten Sinne zu „entfernen“. Dies werden wir später auch betrachten (vgl. Vorgehen II). Vorher wollen wir aber zunächst eine einfachere Methode der „Regularisierung“ im Sinne der Subgroup Discovery betrachten (Vorgehen I). Wir werden genauer analysieren, in welchem Sinne das Hüllensystem aller Begriffsumfänge genau verändert wird und später unsere Analyse mit einer Analyse des Regularisierungsverhaltens bei oben angesprochenem Vorgehen II kontrastieren.

7.2.1 Vorgehen I (Im Stile der Subgroup Discovery): Beschränkung der Description Length

Hier ist die Idee einfach, die Länge der betrachteten subgroup descriptions zu beschränken. Übersetzt in die Sprache der formalen Begriffsanalyse wäre damit konkret gemeint, dass man nur Begriffe betrachtet, die von nicht zu großen Merkmalsmengen (bzw. auch Gegenstandsmengen) erzeugt werden:

Definition 7.1 (K-merkmalserzeugter Begriff, K-gegenstandserzeugter Begriff)
Gegeben ein formaler Kontext $\mathbb{K} = (G, M, I)$ und eine Zahl $K \in \mathbb{N}$ heißt ein formaler Begriff (A, B) des Kontextes $\mathbb{K}$

i) K-merkmalserzeugt, wenn es eine Merkmalsmenge $\tilde{B} \subseteq M$ mit $|\tilde{B}| \leq K$ gibt, die den Begriff (A, B) über
$$(A, B) = (\tilde{B}', \tilde{B}'')$$
erzeugt.

ii) K-gegenstandserzeugt, wenn es eine Gegenstandsmenge $\tilde{A} \subseteq G$ mit $|\tilde{A}| \leq K$ gibt, die den Begriff (A, B) über
$$(A, B) = (\tilde{A}'', \tilde{A}')$$
erzeugt.

Bemerkung 7.1. Natürlich kann ein K-merkmalserzeugter Begriff (bzw. ein K-gegenstandserzeugter Begriff) mehr als K Merkmale (bzw. Gegenstände) beinhalten (bzw. umfassen). Außerdem ist das Mengensystem aller K-merkmalserzeugten Begriffe im Allgemeinen kein Hüllesnystem. (Das gleiche gilt für das Mengensystem aller K-gegenstandserzeugten Begriffe.) Trotzdem kann man eine modifizierte MILP-Formulierung zur Lösung des Optimierungsproblems auf dem reduzierten System aller K-merkmalserzeugten (bzw. auch K-gegenstandserzeugten) Begriffe heranziehen. Dazu müsste man lediglich für den Erzeuger $\tilde{B}$ (bzw. $\tilde{A}$) zusätzliche binäre Variablen einführen.

Nun wollen wir analysieren, wie eine Beschränkung der Description Length sich genau auf die Verkleinerung des Mengensystems auswirkt. In Bezug auf die V.C.-Dimension machen wir die folgenden interessanten Beobachtungen:

Satz 7.2 (V.C.-Dimension und Description Length, vgl. auch [Albano, 2017, Korollar 5.4.5, S.53])
Gegeben sei ein formaler Kontext $\mathbb{K} = (G, M, I)$.

i) Besitzt $\mathbb{K}$ die V.C.-Dimension D, so kann jeder formale Begriff durch D Merkmale bzw. Gegenstände beschrieben werden, d.h., jeder Begriff ist D-merkmalserzeugt und D-gegenstandserzeugt.

ii) Die Umkehrung gilt nicht, d.h., es kann vorkommen, dass jeder formale Begriff D-merkmalserzeugt (bzw. auch D-gegenstandserzeugt) ist, dass der Kontext aber trotzdem eine V.C.-Dimension besitzt, die größer als D ist.

Beweis (Skizze).

i) Habe $\mathbb{K}$ die V.C.-Dimension D und sei (A, B) ein beliebiger formaler Begriff von $\mathbb{K}$. Wir zeigen, dass (A, B) ein D-merkmalserzeugter Begriff ist. (Zu zeigen, dass (A, B) ein D-gegenstandserzeugter Begriff ist, geht analog.) Ist $|B| \leq D$, so sind wir fertig. (Wähle

$\tilde{B} := B$ als Erzeuger mit $|\tilde{B}| \leq D$.) Ist $|B| > D$, so kann B bezogen auf das Hüllensystem aller Begriffsinhalte nicht zerschmetterbar sein, denn die V.C.-Dimension des Hüllesnystsems aller Inhalte (wie auch aller Umfänge) war D. Dies bedeutet, dass B nicht merkmalsimplikationsfrei ist, vergleiche Übung 4, Aufgabe 3. (Da hatten wir nicht Inhalte sondern Umfänge angeschaut, aber die Argumentation ist hier völlig analog.) Es gibt also eine formale Merkmalsimplikation $Y \longrightarrow Z$ mit $\mathbb{Z} \neq \emptyset$, $Y \cap Z = \emptyset$ und $Y \cup Z \subseteq B$, die von allen Gegenständen des Kontexts respektiert wird. Dies bedeutet aber, dass mit $B_1 := Y \subsetneq B$ eine Merkmalsmenge existiert, die echt kleiner ist als B, aber den gleichen formalen Begriff

$$(A, B) = (B_1', B_1'')$$

erzeugt. Ist nun $|B_1| \leq D$, so haben wir gezeigt, dass der Begriff (A, B) ein D-merkmalserzeugter Begriff ist. Andernfalls könnten wir obige Argumentation erneut anwenden und würden nach und nach weitere Merkmalsmengen $B_1 \supsetneq B_2 \supsetneq \ldots B_j$ erhalten, bis schließlich für ein B_j gilt $|B_j| \leq D$, womit wir ebenfalls gezeigt hätten, dass der Begriff (A, B) ein D-merkmalserzeugter Begriff ist. Da (A, B) beliebig war, sind somit alle formalen Begriffe D-merkmalserzeugt.

ii) Um einzusehen, dass die Umkehrung von i) im Allgemeinen nicht gilt, betrachten wir einfach ein Gegenbeispiel:

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
g_1		x			x	x	◯	x
g_2			x		x	◯	x	x
g_3				x	◯	x	x	x

Hier ist jeder Begriff 1-merkmalserzeugt, die V.C.-Dimension ist jedoch 3.

Was bedeutet nun obiger Satz? Sei dazu ein Kontext $\mathbb{K}$ gegeben und sei $\mathcal{S}$ das Hüllensystem aller Begriffsumfänge von $\mathbb{K}$ sowie $\mathcal{S}^K$ das System aller Umfänge von K-merkmalserzeugten Begriffen. Ist nun die V.C.-Dimension von $\mathbb{K}$ kleinergleich K, so würde die Beschränkung der Description Length gar nichts bewirken, d.h., wir hätten

$$\mathcal{S} = \mathcal{S}^K.$$

Ist die V.C.-Dimension größer als K, dann würde die Beschränkung der Description Length das Mengensystem $\mathcal{S}$ unter Umständen verkleinern, konkret aber nur in „Bereichen“ hoher V.C.-Dimension(, was durchaus gewollt ist). Genauer ausgedrückt gilt für beliebige „Bereiche“ $A \subseteq G$ mit „lokaler“ V.C.-Dimension kleinergleich K (im Sinne, dass das projizierte Mengensystem $\mathcal{S}_A = \{B \cap A \mid B \in \mathcal{S}\}$ eine V.C.-Dimension kleinergleich K besitzt), dass

$$\mathcal{S}_A = \left(\mathcal{S}^K\right)_A$$

ist, d.h., in solchen Bereichen wird nichts verändert. Es wird also in der Tat nur in Bereichen hoher lokaler V.C.-Dimension regularisiert.

Leider (?) bedeutet Teil ii) des obigen Satzes, dass nicht notwendigerweise in jedem Bereich hoher lokaler V.C.-Dimension regularisiert wird, dies ist der wesentliche Unterschied zu Vorgehen II. Insgesamt wird also bei Vorgehen I die V.C.-Dimension nicht direkt kontrolliert. Allerdings gilt für die Größe des regularisierten Systems $\mathcal{S}^K$ die Ungleichung

$$|\mathcal{S}^K| \leq \sum_{i=0}^{K} \binom{|M|}{i}.$$

Dies folgt unmittelbar aus der Tatsache, dass es nicht mehr K-merkmalserzeugte Begriffe geben kann, als es subgroup descriptions der Maximallänge K gibt. Betrachtete man nicht K-merkmalserzeugte Begriffe, sondern K-gegenstandserzeugte Begriffe, so erhielte man in analoger Weise für das Mengensystem $\tilde{S}^K$ aller K-gegenstandserzeugten Begriffsumfänge die Ungleichung

$$|\tilde{\mathcal{S}}^K| \leq \sum_{i=0}^{K} \binom{|G|}{i},$$

was erstaunlicherweise genau die gleiche Abschätzung wie bei einer V.C.-Analyse und V.C.-Dimension K ergäbe.(Allerdings scheint mir in der klassischen Subgroup Discovery die Betrachtung von K-gegenstandserzeugten Begriffen anstelle von K-merkmalserzeugten Begriffen nicht üblich zu sein.) Dies führt natürlich unmittelbar zu folgender

Frage: Sollte man eher die V.C.-Dimension oder eher direkt die Größe des unterliegenden Mengensystems kontrollieren?

Antwort: Mir unbekannt.

Natürlich ist dies wohl keine wohlgestellte Frage, außerdem hängen die V.C.-Dimension und die Größe eines Mengensystems ja recht unmittelabr miteinander zusammen, es gilt doch

$$2^{VC(\mathcal{S})} \leq |\mathcal{S}| \leq \sum_{i=0}^{VC(\mathcal{S})} \binom{|V|}{i}.$$

7.2.2 Vorgehen II (Im Stile der Formalen Begriffsanalyse): Identifikation großer Kontranominalskalen

Kommen wir nun zu einer zweiten Möglichkeit der Regularisierung, die allerdings rein computational etwas aufwendiger ist. Wir wissen, dass die zerschmetterbaren Mengen genau durch die Kontranominalskalen gegeben sind. Außerdem kann man für einen gegebenen Kontext $\mathbb{K} = (G, M, I)$ große Kontranominalskalen mit Hilfe eines MILP-Ansatzes identifizieren, was computational aber schon recht aufwendig ist, insbesondere kann es ja auch sehr viele Kontranominalskalen der Größe der V.C.-Dimension von $\mathbb{K}$ geben. Davon einmal abgesehen stellt sich die Frage, was man tun soll, nachdem man alle Kontranominalskalen der Größe $K = VC(\mathbb{K})$ (oder auch größergleich K mit $K < VC(\mathbb{K})$) identifiziert hat. Eine Möglichkeit, die wir nun kurz analysieren wollen, bestünde darin, diejenigen **Gegenstände**, die zu den identifizierten Kontranominalskalen gehören, zunächst für die Erzeugung des betrachteten Hüllensystems von Begriffsinhalten nicht in Betracht zu ziehen. (Bei der schlussendlichen Berechnung der Supremumstatistik werden diese Gegenstände dann natürlich wieder betrachtet.) Sei dazu sh_K die Menge aller zerschmetterbaren Gegenstandsmengen der Kardinalität größergleich K. Sei weiter $N := \bigcup sh_K$ die Menge aller Gegenstände, die zu zerschmetterbaren Mengen der Kardinalität größergleich K gehören. Betrachte dann den verkleinerten Kontext

$$\mathbb{K}_{\backslash N} := (G \backslash N, M, I \cap G \backslash N \times M)$$

und definiere nun das Hüllensystem

$$\mathcal{T}^K := \{\{g \in G \mid \forall m \in B : gIm\} \mid B \text{ Begriffsinhalt von } \mathbb{K}_{\setminus N}\}$$

aller Begriffsumfänge des ursprünglichen Kontextes $\mathbb{K}$, die von Begriffsinhalten des kleineren Kontextes $\mathbb{K}_{\setminus N}$ erzeugt werden. Dann gilt folgender

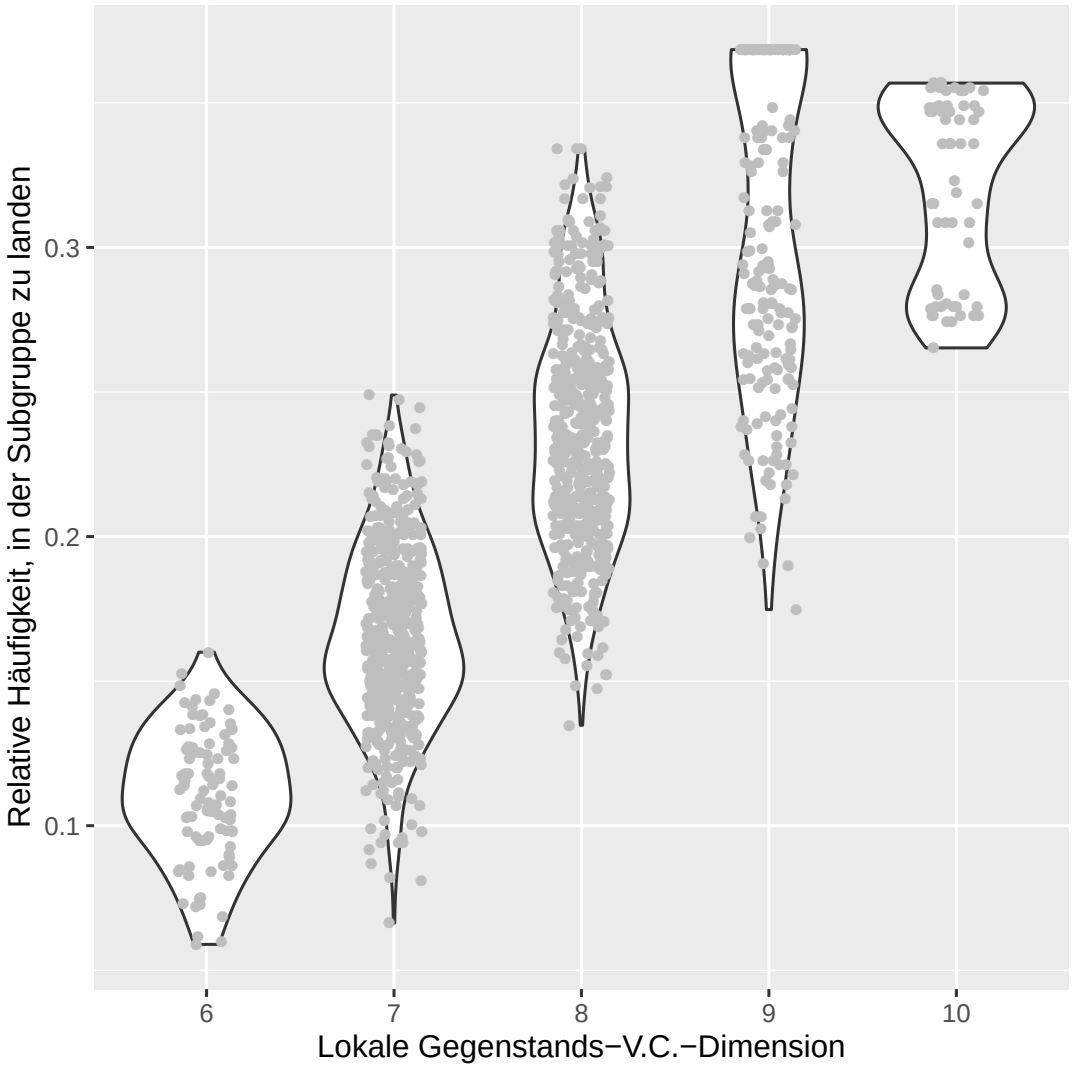
Satz 7.3 (V.C.-Dimension bei lokaler Entfernung großer Kontranominalskalen)
Es gilt:

i) $\mathcal{T}^K \subseteq \mathcal{S}$

ii) Die V.C.-Dimension von $\mathcal{T}^K$ ist maximal K.

iii) Ist die lokale V.C.-Dimension von $\mathcal{S}$ in einem Bereich $A \subseteq G$ kleinergleich K(, d.h., $\mathcal{S}_A$ hat V.C.-Dimension kleinergleich K), so gilt $\mathcal{S}_A = \left(\mathcal{T}^K\right)_A$.

Obiger Satz sagt also genau das aus, was man sich wünschen könnte: Es wird genau in Bereichen hoher lokaler V.C.-Dimension regularisiert. Ob eine solche Form der Regularisierung wirklich wünschenswert ist, ist natürlich eine sehr schwierige Frage. Ein Aspekt, an den man bezogen auf die V.C.-Dimension denken könnte (, der aber wohl so nicht direkt als Versuch einer Antwort herhalten kann ...), soll nun abschließend an dem folgenden kleinen Beispiel kurz angerissen werden: Wir hatten oben überlegt, dass gerade zerschmetterbare Mengen bezogen auf das Verhalten der Teststatistik ungünstig sind. Dies könnte zu der Spekulation verführen, dass nicht nur große Mengensysteme $\mathcal{S}$ das Verhalten der Teststatistik (im Sinne des Beweises der hinreichenden Bedingung für Konvergenz) negativ beeinflussen, sondern dass vielmehr insbesondere die zerschmetterbaren Mengen einen viel negativeren Einfluss das Verhalten der Teststatistik (im Sinne der notwendigen Bedingung) besitzen. Um sich hier einen (eventuell trügerischen?) ersten Einblick zu verschaffen, betrachten wir jetzt nochmal eine Art lokale Variante der V.C.-Dimension: Für einen Gegenstand $g \in G$ definiere die lokale Gegengstands-V.C.-Dimension als die maximale Kardinalität einer zerschmetterbaren Menge, die den Gegenstand g enthält. Die folgende Graphik zeigt nun beispielhaft eine Simulation für das Allbus-Datenbeispiel zur Subgroup Discovery (bei nominaler Skalierung). Aufgetragen ist auf der x-Achse die lokale Gegenstands-V.C.-Dimension für alle 1354 Einheiten. Auf der y-Achse ist die relative Häufigkeit dafür aufgetragen, dass die betrachtete Einheit in der berechneten Subgruppe mit dem höchsten Wert der Piatetsky-Shapiro-Qualitätsfunktion landet, und zwar unter einer Simulation unter der Nullhypothese, dass die Zielvariable stochastisch unabhängig ist von allem Anderen. Unter der Annahme, dass die Zielvariable völlig bedeutungslos ist, könnte man sich erwarten, dass nicht bestimmte statistische Einheiten systematisch häufiger in der Subgruppe landen. (Ob ein solches eventuell wünschenswertes Verhalten unter H_0 in ausreichender Weise erreichbar ist, scheint eine eher schwierige Frage zu sein.) Man kann in der Graphik klar erkennen, dass die relative Häufigkeit dafür, in der Subgruppe zu landen von der Tendenz her klar mit der lokalen Gegenstands-V.C.-Dimension steigt. (Beispielsweise der Korrelationskoeffizient nach Spearman ist hier etwa 0.84.) Dies könnte man jetzt vorläufig so interpretieren, dass die entdeckte Subgruppe unter H_0 systematisch „in Richtung von Bereichen hoher V.C.-Dimension verzerrt“ ist. Ob die V.C.-Dimension hier die angemessenste Charakteristik ist, um „das Phänomen“ zu beschreiben, bleibt natürlich fraglich. Wollen wir es damit mit Spekulationen auf sich beruhen lassen.....


Relative Häufigkeit, in der Subgruppe zu landen
0.1
0.2
0.3
6
7
8
9
10
Lokale Gegenstands−V.C.−Dimension

"... and it is probable that there is some secret here which remains to be discovered." [Peirce]

Literatur

A. L. J. H. Albano. *Polynomial Growth of Concept Lattices, Canonical Bases and Generators: Extremal Set Theory in Formal Concept Analysis*. PhD thesis, Saechsische Landesbibliothek-Staats-und Universitaetsbibliothek Dresden, 2017.

N. A. L. Ayastuy. *Dinámica espacio-temporal de un bosque secundario en el Parque Natural de Urkiola (Bizkaia): tesis doctoral*. PhD thesis, 2008.

A. Beutelspacher. *Einführung in die endliche Geometrie: 1: Blockpläne.–247 s.* Bibliographisches Institut, 1982.

B. A. Davey and H. A. Priestley. *Introduction to lattices and order*. Cambridge university press, 2002.

M. de la Cruz Rot. *Metodos para analizar datos puntuales.*, chapter 3, pages 76–127. Asociacion Espanola de Ecologia Terrestre, Universidad Rey Juan Carlos and Caja de Ahorros del Mediterraneo, 2008.

G. Fasano and A. Franceschini. A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 03 1987. ISSN 0035-8711. doi: 10.1093/mnras/225.1.155. URL `https://doi.org/10.1093/mnras/225.1.155`.

M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.

B. Ganter. Formale Begriffsanalyse. *Wissenschaftliche Zeitschrift der Technischen Universität Dresden*, 45(6):8–13, 1996.

B. Ganter. *Diskrete Mathematik: Geordnete Mengen*. Springer, 2013.

M. D. Lee, M. Steyvers, and B. Miller. A cognitive model for aggregating people's rankings. *PloS one*, 9(5):e96431, 2014.

J. A. Major and J. J. Mangano. Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems*, 4(1):39–52, Jan 1995. ISSN 1573-7675. doi: 10.1007/BF00962821. URL `https://doi.org/10.1007/BF00962821`.

P. Muldowney, K. Ostaszewski, and W. Wojdowski. The darth vader rule. *Tatra Mountains Mathematical Publications*, 52(1):53–63, 2012.

A. Pajor. *Sous-espaces LN/L des espaces de Banach*. Number 16. Hermann, 1985.

J. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 1983.

C. S. Peirce. *Über die Klarheit unserer Gedanken*. Popular Science Monthly 12, 1878. Deutsche Übersetzung: Klaus Oehler 1968.

G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–238, 1991. URL `https://ci.nii.ac.jp/naid/10000000985/en/`.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145 – 147, 1972. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(72)90019-2. URL `http://www.sciencedirect.com/science/article/pii/0097316572900192`.

G. Schollmeyer. Application of lower quantiles for complete lattices to ranking data: analyzing outlyingness of preference orderings, 2017. URL `https://epub.ub.uni-muenchen.de/40452/1/TR_208.pdf`.

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

S. E. Syrjala. A statistical test for a difference between the spatial distributions of two populations: Ecological archives e077-001. *Ecology*, 77(1):75–80, 1996.

V. Vapnik. Interview with Vladimir Vapnik: Statistical Learning. MIT Artificial Intelligence (AI) Podcast, 2018. URL `https://www.youtube.com/watch?v=STFcvzoxVw4`.

V. N. Vapnik. *Estimation of dependences based on empirical data. Empirical inference science: Afterword of 2006 / Vladimir Vapnik*. Springer, New York, NY and [Heidelberg], 2006. ISBN 0387308652.

V. N. Vapnik and A. Y. Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences, 1968.

J. Wollbold. Attribute exploration of gene regulatory processes. *arXiv preprint arXiv:1204.1995*, 2012.

S. Wrobel. Inductive logic programming for knowledge discovery in databases. In *Relational data mining*, pages 74–101. Springer, 2001.

A Notation

$\mathbb{N} := \{0, 1, 2, \ldots\}$	Menge der natürlichen Zahlen
$\mathbb{N}_+$	Menge der positiven natürlichen Zahlen
2^Ω	Menge aller Teilmengen der Menge Ω
A^c	Komplement der Menge A
$\mathbb{1}_A$	Indikatorfunktion der Menge A
$C[a, b]$	Menge aller stetigen, reellwertigen Funktionen auf dem Intervall $[a, b]$
id_T	identische Abbildung auf der Menge T
B^A	Menge aller Abbildungen von A nach B
$\Psi \circ \Phi$	Komposition der Abbildungen Φ und Ψ
$\leq_{SD}$	stochastische Dominanz erster Ordnung
$\downarrow x := \{y \mid y \leq x\}$	Vorbereich des Elements x bezüglich $\leq$
$\downarrow S := \{y \mid \exists x \in S : y \leq x\}$	Vorbereich der Menge S bezüglich $\leq$
$\uparrow x := \{y \mid y \geq x\}$	Nachbereich des Elements x bezüglich $\leq$
$\uparrow S := \{y \mid \exists x \in S : y \geq x\}$	Nachbereich der Menge S bezüglich $\leq$
$\mathfrak{O}((V, \leq))$	Menge aller Oberhalbmengen von $(V, \leq)$
$\lessdot$	Nachbarschaftsrelation zur Relation $\leq$
$\bot$	kleinstes Element eines vollständigen Verbandes
$\top$	größtes Element eines vollständigen Verbandes
$\bigvee_{t \in T} x_t$	Supremum der Menge $\{x_t \mid t \in T\}$
$\bigwedge_{t \in T} x_t$	Infimum der Menge $\{x_t \mid t \in T\}$
$F_X,\ F$	Verteilungfunktion der Zufallsvariablen X
$\hat{F}_x,\ \hat{F}$	empirische Verteilungsfunktion des Datenvektors x
F_n	empirische Verteilungsfunktion der ersten n Datenpunkte eines i.i.d.-samples
$\hat{\mathbb{P}}_x,\ \hat{\mathbb{P}}$	empirisches Wahrscheinlichkeitsmaß zum Datenvektor x
$\mathbb{P}_n$	empirisches Wahrscheinlichkeitsmaß der ersten n Datenpunkte eines i.i.d.-Samples
$\lvert T \rvert$	Anzahl der Elemente der Menge T
$\mathrm{co}\,(T)$	konvexe Hülle der Menge T
$A + z$	Komplexnotation für $A + z := \{a + z \mid a \in A\}$
$\forall$	Allquantor
$\exists$	Existenzquantor
$\&$	logisches und
$\sqsubseteq \supseteq \leq$	Die Relation $\sqsubseteq$ enthält die Relationd $\leq$, d.h., es gilt $(x, y) \in \leq \Longrightarrow (x, y) \in \sqsubseteq$
$im(X) := \{X(\omega) \mid \omega \in \Omega\}$	Bild der Funktion (Zufallsvariable) X
$(x)_+ := \max(0, x)$	Positivteil der reellen Zahl x

Index