Estimating the size of meet-distributive concept lattices. With an application to statistical inference via Vapnik Chervonenkis theory ... ähm ... Rademacher complexity ähm ... the simple union bound.

PART (I+) II

... These thoughts are due to my impatience while waiting for Gurobi to finish the calculations. ...

1. A motivating example & some comments (on statistical learning theory/statistics)

2. Further examples

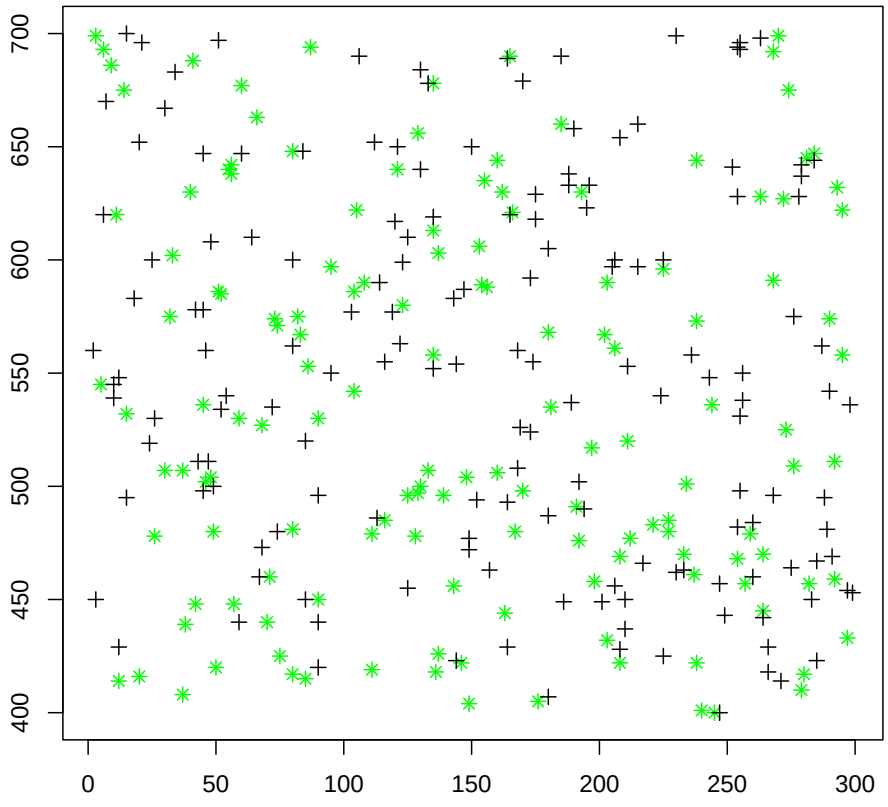
3. Reminder: Formal concept analysis

4. Counting concepts

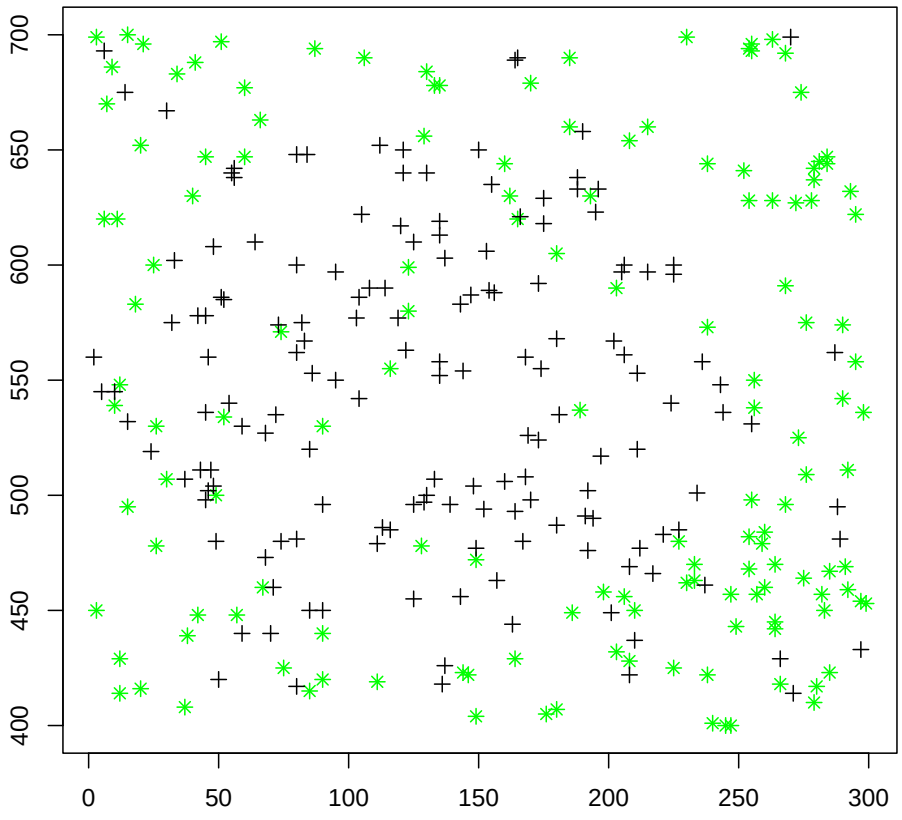
A motivating example & some comments (on statistical learning theory/statistics)

A motivating example: Locations of plants in a gypsophylous plant community in central Spain

A: H_0?



B: H_0?



H_0**:** The spatial distribution of *green* and *black* plants is the same.

A motivating example: Kolmogorov-Smirnov type test for differences between spatial distributions of two sub-populations

$$D_n := \sup_{A \in \mathcal{S}} \left| \hat{P}_1(A) - \hat{P}_2(A) \right|$$

with

$$\mathcal{S} := \{ A \cap X_{obs} \mid A \subseteq \mathbb{R}^2, A \text{ convex} \} \text{ (a } \textbf{finite!} \text{ closure system)}$$

X_{obs} : the set of all observed points in $\mathbb{R}^d$

$\hat{P}_1$: empirical law of (sample of) subpopulation 1 of size n_1

$\hat{P}_2$: empirical law of (sample of) subpopulation 2 of size n_2

$$n := min\{n_1, n_2\}$$

additionally set $N := n_1 + n_2; \quad m := max\{n_1, n_2\};$

$$D_n^+ := \sup_{A \in \mathcal{S}} \hat{P}_1(A) - \hat{P}_2(A); \quad D_n^- := \inf_{A \in \mathcal{S}} \hat{P}_1(A) - \hat{P}_2(A)$$

Motivation/Question of interest

- $H_0 : P_1 = P_2 \quad vs. \quad H_1 : P_1 \neq P_2$
- If D_n is large enough, reject H_0
- How to assess statistical significance (i.e., how large is large?)?:
  - A) permutation test (computationally demanding)
  - B) VC analysis (very conservative)
  - C) Rademacher type analysis (computationally as demanding as, and in terms of conservativeness not better than A)
  - **D) simple union-bound analysis including estimation of $|\mathcal{S}|$.** (Conservativeness lies between that of A and B, computationally attractive in many cases)

Under i.i.d. (?) sampling, under the null (with fixed X_{obs} for ease of exposition, otherwise conditional analysis)

$$P(D_n > \varepsilon) \leq \underbrace{\underbrace{\frac{9}{2} \cdot \frac{N^{\hbar}}{\hbar!}}_{=:C_{\mathcal{VC}}}}_{=:\alpha(\varepsilon)} \cdot \exp\{-n \cdot \varepsilon^2\} \qquad (?, \text{p. } 172)$$

with $N := n_1 + n_2$ (and assuming $n_1 = n_2$) and $\hbar$ the VC dimension of the fixed and finite family $\mathcal{S}$.

A better bound for the case of finite $\mathcal{S}$

$$P(D_n > \varepsilon) \leq \underbrace{4 \cdot |\mathcal{S}| \cdot \exp\left[-\frac{n \cdot N}{2 \cdot (m+1)} \cdot \varepsilon^2\right]}_{=:\alpha_{ub}(|\mathcal{S}|,n,m,\varepsilon)=:\Gamma}$$

or for $n = n_1 = n_2$:

$$P(D_n > \varepsilon) \leq 4 \cdot |\mathcal{S}| \cdot \exp\left[-n \cdot \frac{n}{n+1}\varepsilon^2\right]$$

$$\approx 4 \cdot |\mathcal{S}| \cdot \exp\left[-n \cdot \varepsilon^2\right]$$

(triangle inequality plus union bound plus **?**)

Note: Up to subtleties this is the same result as what follows from the analysis of the growth function.

"Reminder" I

Hoeffding: $P(P(A) - \hat{P}^n(A) > \varepsilon) \leq \exp\left[-2n\varepsilon^2\right]$
(with fixed event A and i.i.d sample of size n)

Serfling: $P(P(A) - \hat{P}^n(A) > \varepsilon) \leq \exp\left[-2n\varepsilon^2/(1 - f_n^*)\right]$
(for sampling without replacement from a population of size N and sample of size n and $f_n^* := (n-1)/N$)

Union bound:

$$P\left(\bigcup_{k=1}^{L} A_k\right) \leq \sum_{k=1}^{L} P(A_k) \leq L \cdot c$$

with $c := \sup_{k \in \{1,\ldots,L\}} P(A_k)$

“Reminder” II

Triangle inequality (here):

$$P(|\hat{P}_1(A) - \hat{P}_2(A)| > \varepsilon) \quad \leq \quad P(|\hat{P}_1(A) - P(A)| > \tfrac{\varepsilon}{2}) + P(|\hat{P}_2(A) - P(A)| > \tfrac{\varepsilon}{2})$$

Reason:

$$\text{“}|\hat{P}_1(A) - \hat{P}_2(A)| > \varepsilon\text{”} \quad \subseteq \quad \text{“}|\hat{P}_1(A) - P(A)| > \tfrac{\varepsilon}{2}\text{”} \cup \text{“}|\hat{P}_2(A) - P(A)| > \tfrac{\varepsilon}{2}\text{”}$$

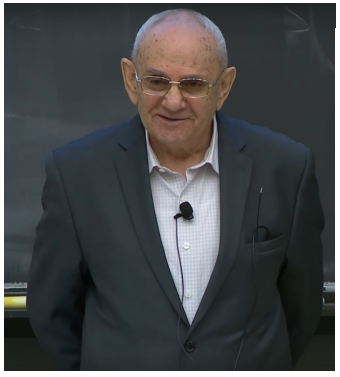
Comments

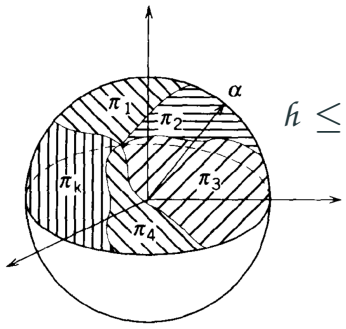
"The reality is that the VC line of analysis leads to a very loose bound..."

"... Second, although the bound is loose, it tends to be equally loose for different learning models, and hence is useful for comparing the generalization performance of these models. This is an observation from practical experience, not a mathematical statement. "

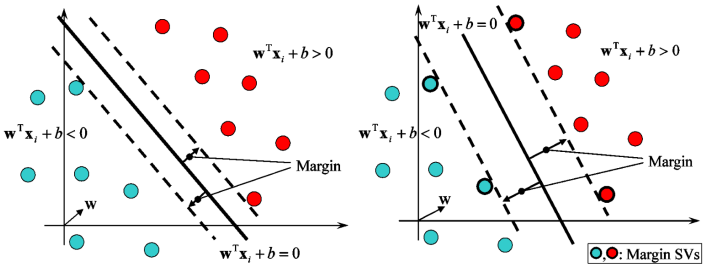
"Thus, the VC bound can be used as a guideline for generalization, relatively if not absolutely. " (**?**)

Comments





$$\hbar \leq min\left(\left[D^2A^2\right], n\right) + 1$$



(a) Small margin (b) Optimized large margin

"There is a mathematical setting. When I came to [the] United State[s] in 1990 first, people did not know VC theory, they did not know statistical learning theory. In Russia, it was published two monographs, our monographs, but in America they did not know. Then, they learned 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 .... VC dimension you can prove some theorems ..." [Vapnik 2018]

Comments


Rademacher Complexity:

"Unlike the VC dimension based bounds, which were distribution independent, the Rademacher complexity bounds depend on the training set distribution, and thus can give better bounds for specific input distributions. Furthermore, the Rademacher complexity can, **in principle**, *be estimated from the training set, allowing for strong bounds derived from a sample itself."* (**?**)

$$\mathbb{E}\left(\sup_{A\in\mathcal{S}}|P(A)-\hat{P}^N(A)|\right)\leq 2\mathcal{R}$$

with

$$\mathcal{R}:=\mathbb{E}\left[\sup_{A\in\mathcal{S}}\frac{1}{N}\sum_{x\in X_{obs}}^{N}\sigma_i\mathbb{1}_A(x)\right]$$

with $\sigma_1,\ldots,\sigma_N$ i.i.d. Rademacher distributed (i.e. $P(\sigma_i=-1)=P(\sigma_i=1)=0.5$.)

But: Directly estimating Rademacher complexity is as computational expensive as doing a permutation test! (In our cases, often the bottleneck is computing the supremum type statistic.) Other techniques like Massart's lemma would require to estimate $|\mathcal{S}|$.

Aims within this presentation

- Estimate $|\mathcal{S}|$ to assess statistical significance of a distributional test
- Estimating $|\mathcal{S}|$ is also a question of its own interest and with further applications, e.g.:
  - Quick check if computation of a large concept lattice is computationally feasible at all (c.f., also (**?**))
  - Uniform regularization by locally controlling $|S_j|$ to regularize $S := \bigcup_{j \in J} S_j$ (e.g. in the context of star-shaped subgroup discovery)
  - ...

Aims within this presentation

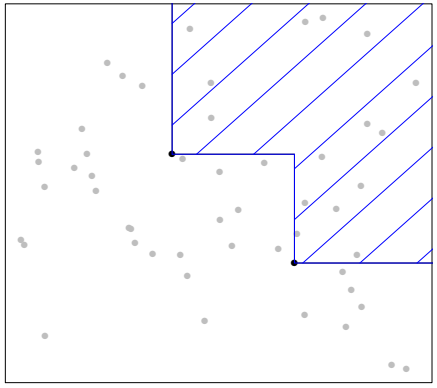
- Works for meet-distributive closure systems/concept lattices
- In the non-meet-distributive case one might work with meet-distributive (upper) approximations
- Meet-distributive approximations are of its own interest (e.g., in the context of data depth within FCA or in the context of robustness and FCA)

Further examples

Further examples of meet-distributive closure systems

upsets

U upset iff $\forall a \leq b$:
$a \in U \Longrightarrow b \in U$.

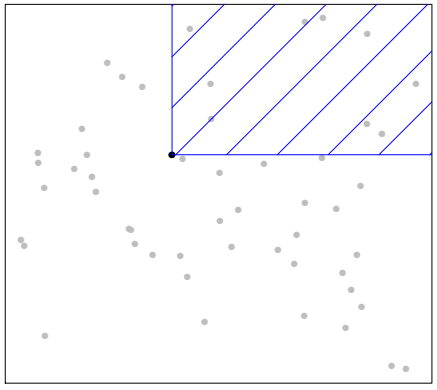


Application example: multivariate poverty-/ inequality analysis (3 dimensions income, education and health)

VC dimension = maximal number of corners

principal filters

F principal filter if
$F = \{y \in V \mid y \geq c\}$ for some $c \in V$.

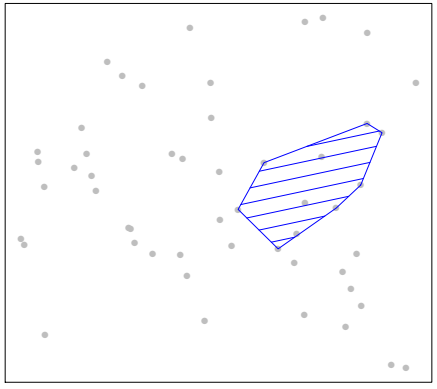


possible applications: multidimensional K.-S.-tests, e.g. item response theory (item impact or DIF)

VC dimension = 1
FALSCH!!

convex sets

C convex if $\forall x, y, z \in V, w \in conv(\{x, y, z\})$:
$x, y, z \in C \Longrightarrow w \in C$



possible applications: spatial statistics

VC dimension = ∞ (or maximal number of extreme points)

Further examples of meet-distributive closure systems

- Subgroup discovery with only interordinally scaled variables without ties.
- Apposition[1] of many meet-distributive contexts (e.g., spatial data (without ties) plus one or more numeric variable(s) (without ties)).
- Local rings of sets ... are locally meet-distributive (after factorizing over non-antisymmetries)

[1] In the sense of combining many meet-distributive contexts to one context by using all attributes from every context.

Example: convex sets

$N = 300$

$|\mathcal{S}| \approx 5.7 \cdot 10^{15}$

t-CI: $\left[0; 1.9 \cdot 10^{16}\right])$

abc-CI: $\left[10^{15}; 1.7 \cdot 10^{17}\right]$

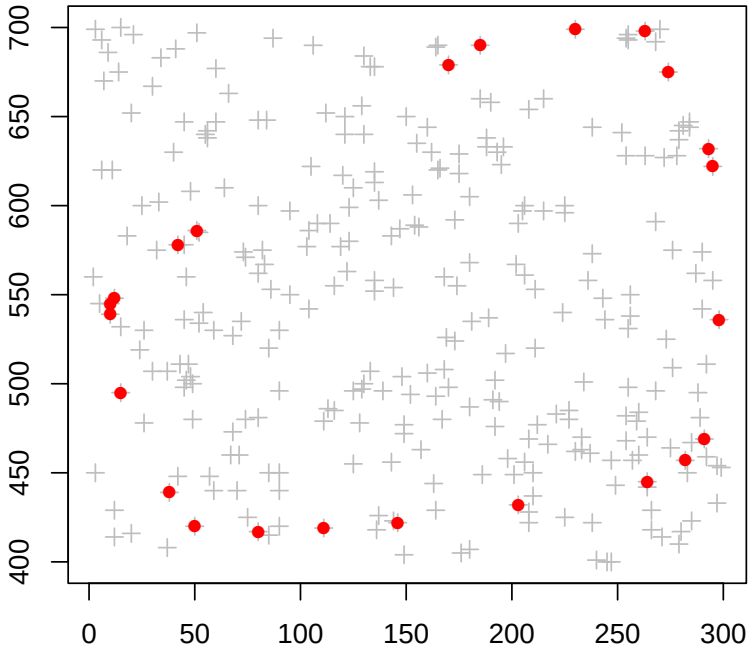
confidence level: $1 - 10^{-10}$

based on sample of size 540

computation time ca $6h$

$h \geq 27$ (GUROBI: out of memory)

$\mathcal{C}_{\mathcal{VC}} \geq 3.2 \cdot 10^{39}$



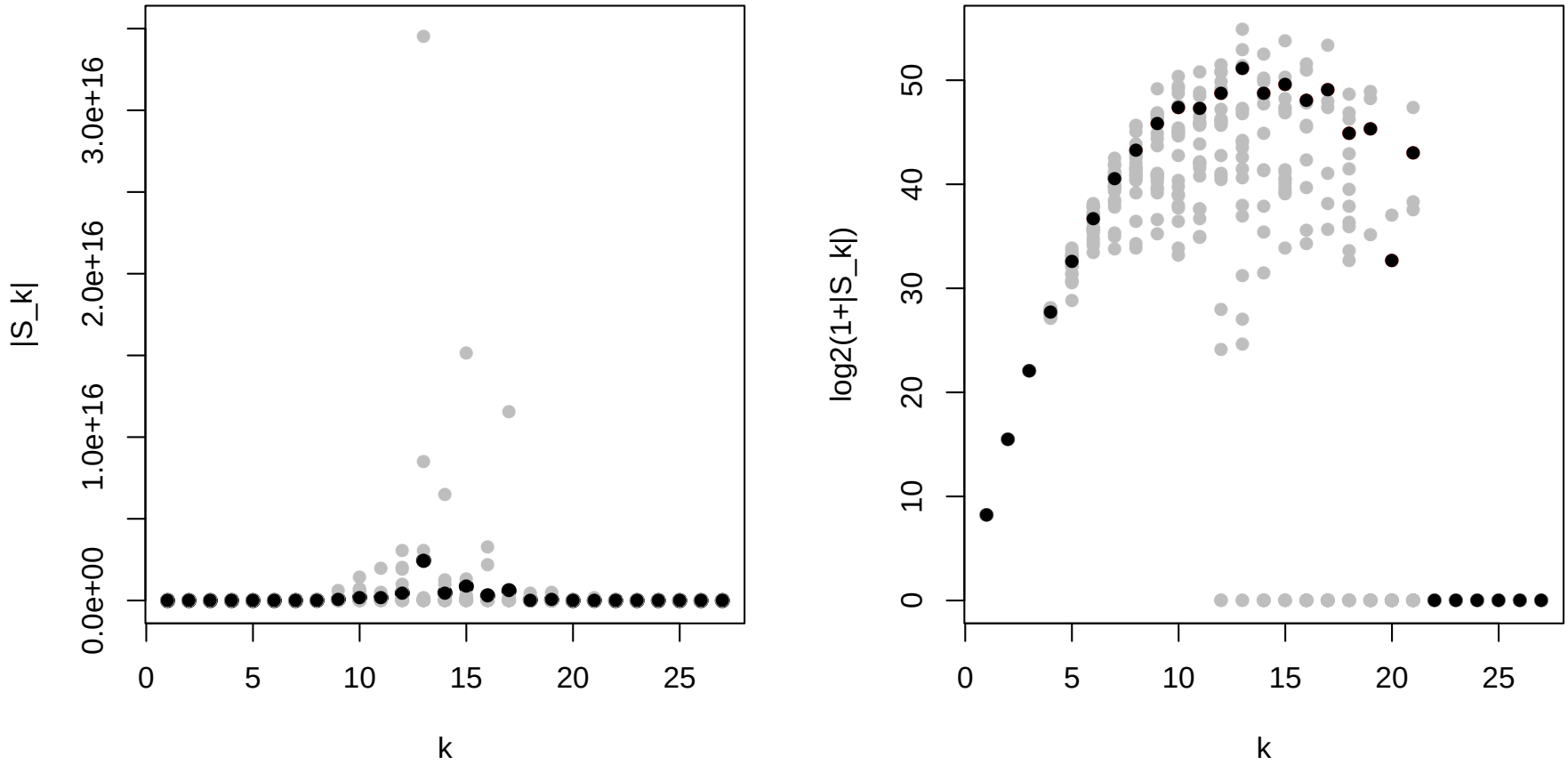


Figure 1: Estimated number N_k of convex sets with k extreme points (left linear, right logarithmic display

Example: convex sets: (some) H_1

$n = 143, \quad m = 157, N = 300$

$D_n^{obs} \approx 0.728$

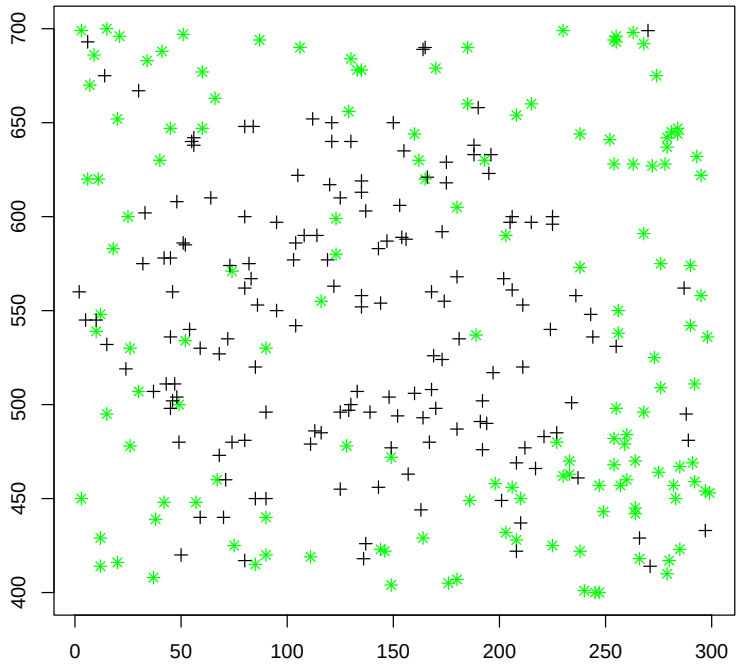
$P_{H_0}(D_n > D_n^{obs}) \leq \Gamma \approx 1.2 \cdot 10^{-15}$

abc-CI for Γ: $\left[2.2 \cdot 10^{-16}; 3.7 \cdot 10^{-14}\right]$

runtime GUROBI (with tricks): $129s$

$\hbar \geq 27$ (out of memory)

VC bound useless (too loose)



Example: convex sets: H_0

$n = 143, \quad m = 157, N = 300$

$D_n^{obs} \in [0.23;\ 0.54]$

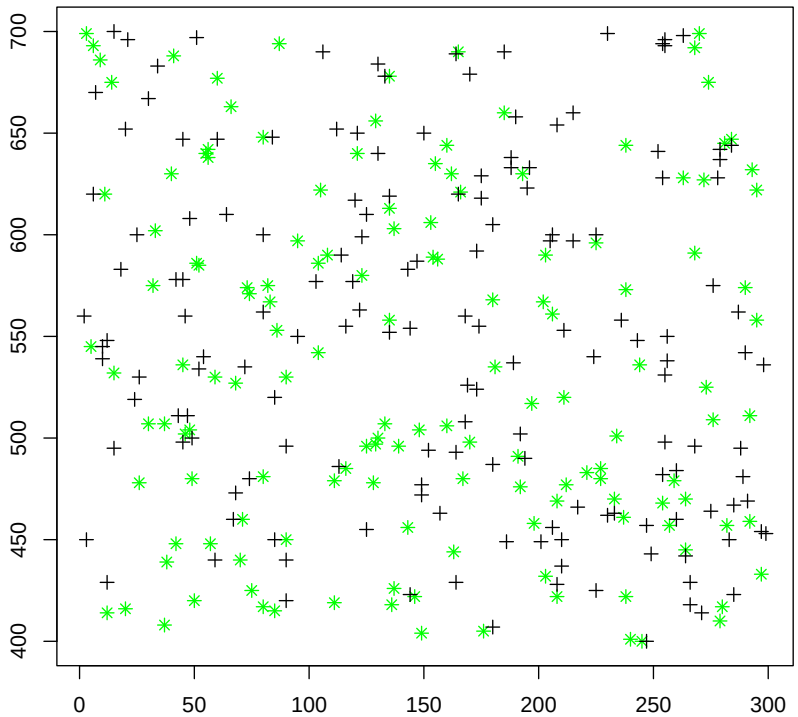
$P_{H_0}(D_n > D_n^{obs}) \leq 5.2 \cdot 10^{14}$?

runtime GUROBI (despite tricks): $\geq 5687s$

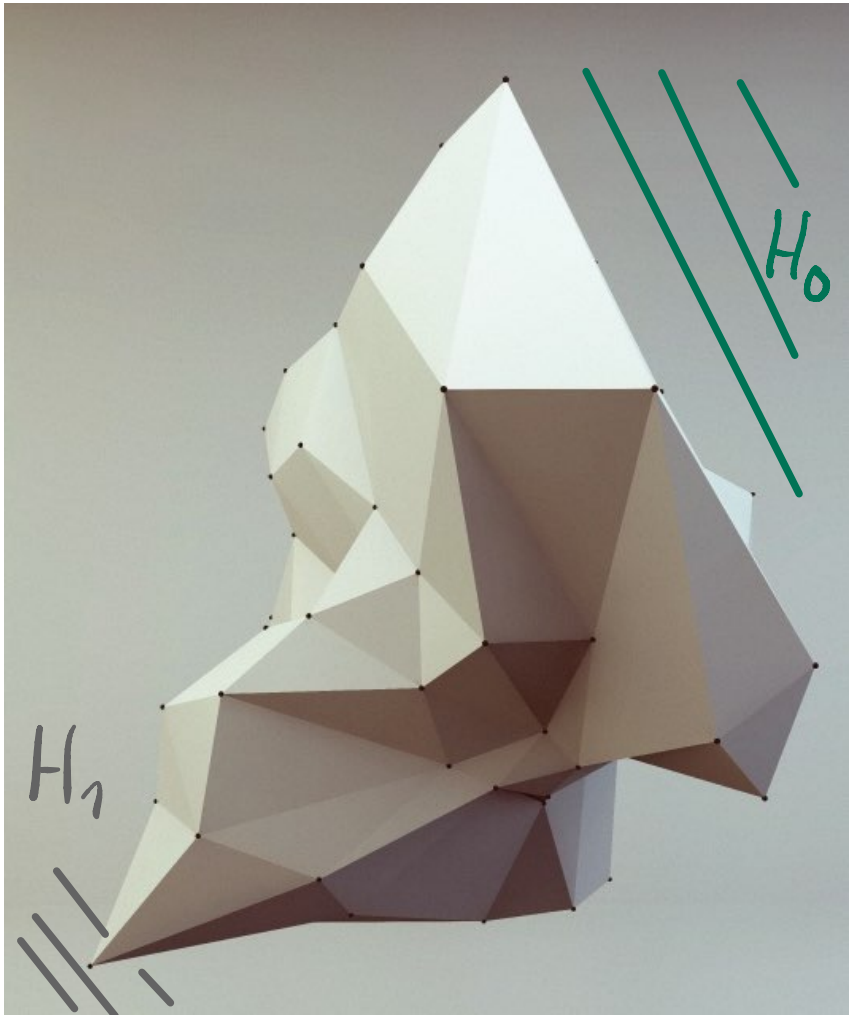
(out of memory)

or: $217s$ for deciding $D_n^{H_0} < D_n^{obs}$

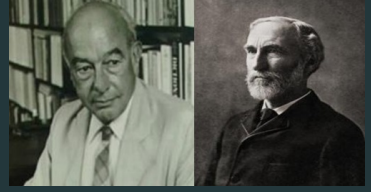
for one resample



Side remark



Comment on Duhem-Quine / i.i.d.



- What if (especially in the spatial case) the sample is not i.i.d. under the null?
- For example for the case of $(r-1)$-dependent random variables[2], Hoeffding's inequality (**?**) gives

$$P(|P(A) - \hat{P}^N(A)| \geq \varepsilon) \leq 2 \cdot \exp\left[-2\,(n/r)\,\varepsilon^2\right]$$

and I presume that a similar statement is valid for **?**. Together with the union bound this would give

$$P(D_n > \varepsilon) \lessapprox 4 \cdot |\mathcal{S}| \cdot \exp\left[-(n/r) \cdot \varepsilon^2\right],$$

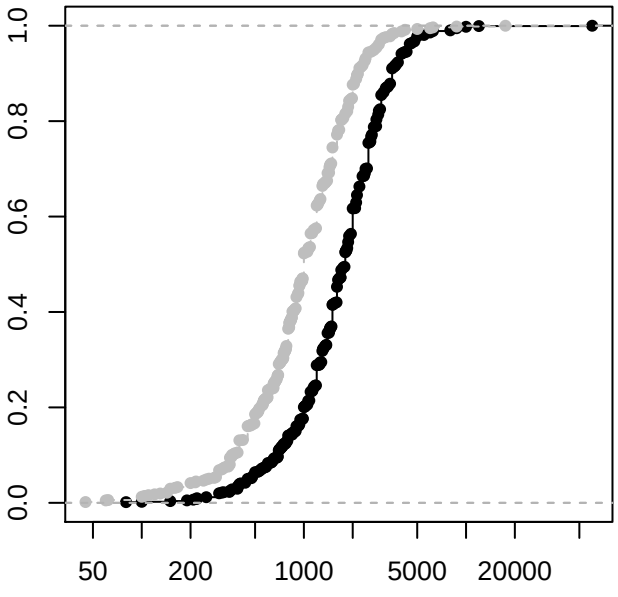
and n/r could therefore be vaguely interpreted as an *effective sample size*.

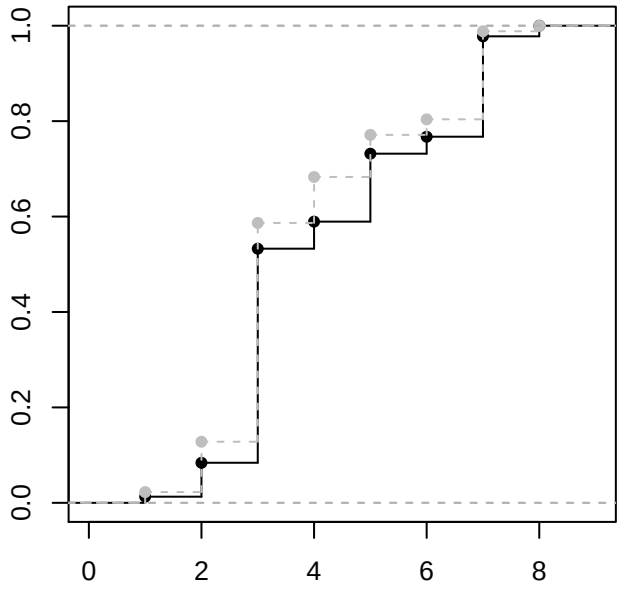
[2] A random vector $(X_1, \ldots, X_n)$ is called $(r-1)$ dependent if for $j - i \geq r$ the random vectors $(X_1, \ldots, X_i)$ and $(X_j, \ldots, X_N)$ are independent.

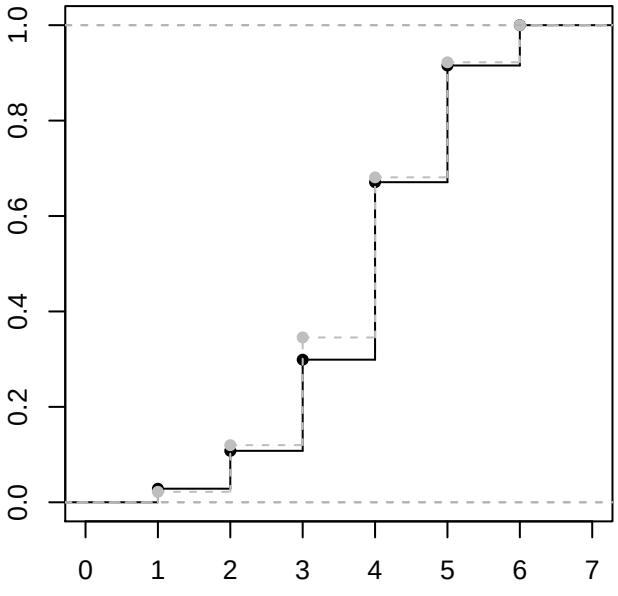
Another application example (upsets)

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

Marginal analysis

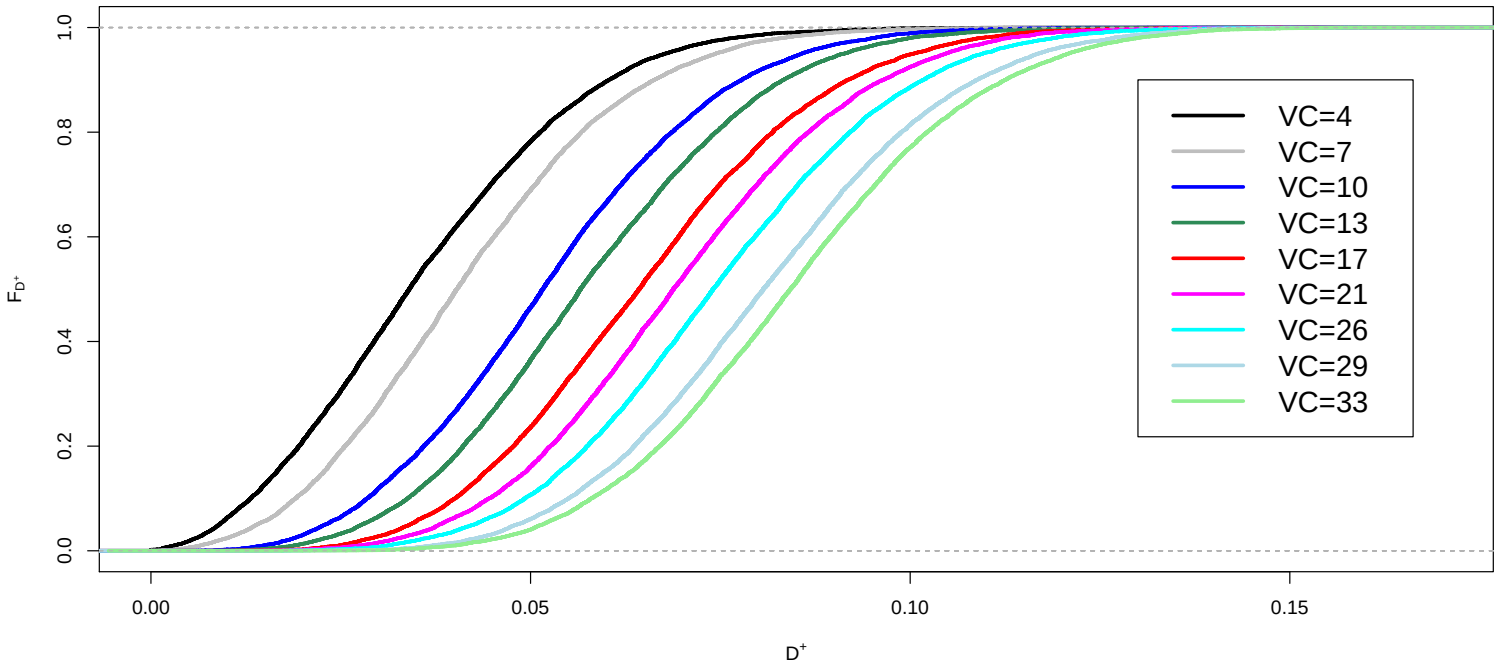






Joint analysis

- $\hbar = 33$ (number of upsets $\in [10^{10}, 10^{60}]$, dual simplex algorithm took less than a second).
- $D_n = D_n^+ \approx 0.36$. ($D_n^- \approx -1.2\%$, female subgroup almost stochastically smaller than male subgroup).
- Value of D_n^+ significantly positive according to a permutation test. (D_n^- not significantly different from zero.)



Estimating $|\mathcal{S}|$

- $\hbar = 33$, therefore we have $|A| \leq 33$ for every minimal generator $A \in \mathit{mingen}(\mathcal{S})$ (see later).
- estimate of 3300 sampled minimal generators (took ca. 1.4 min time):

$$|\mathcal{S}| \approx 7 \cdot 10^{20}$$
$$CI : [0; 1.9 \cdot 10^{21}] \text{ (confidence level: } 1 - 10^{-10})$$
$$\Gamma \approx 1.9 \cdot 10^{-16} \quad (CI : [0; 5.2 \cdot 10^{-16}]$$

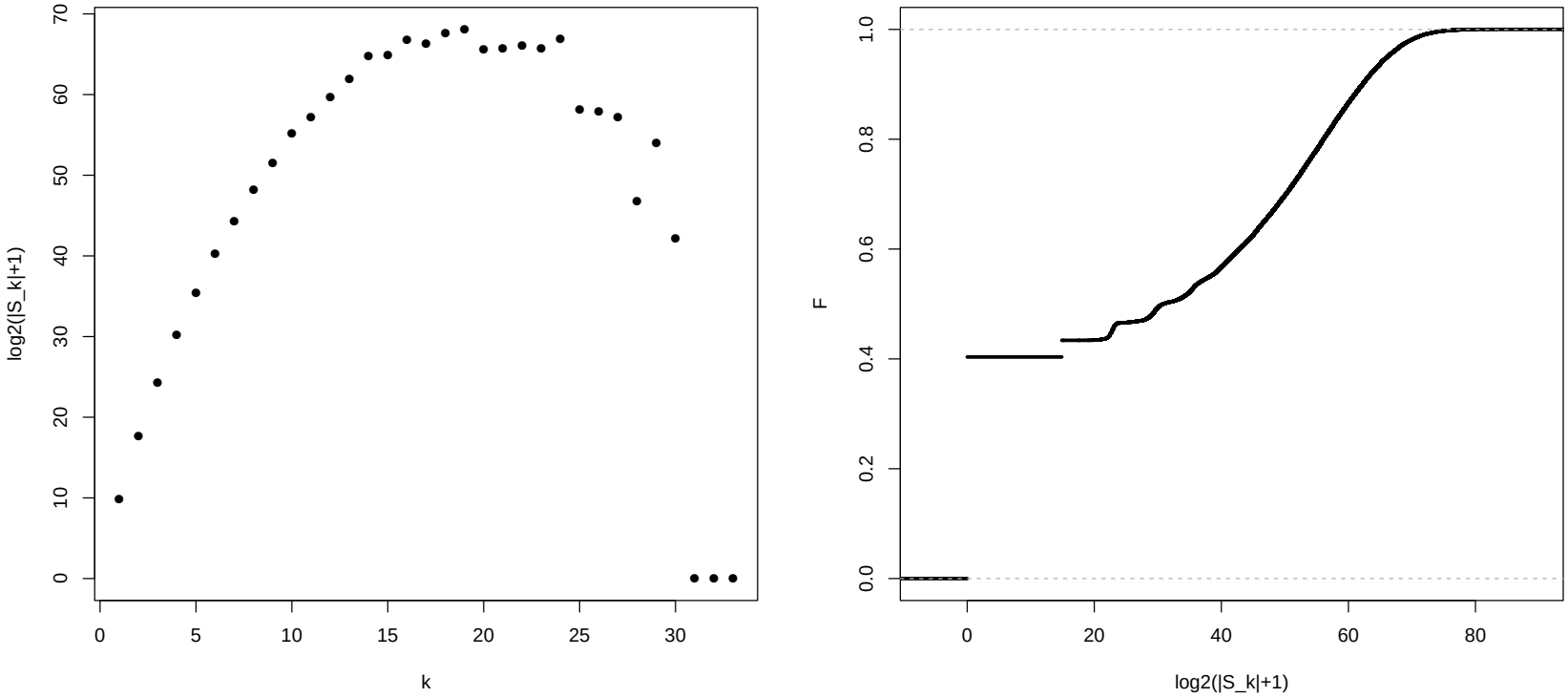
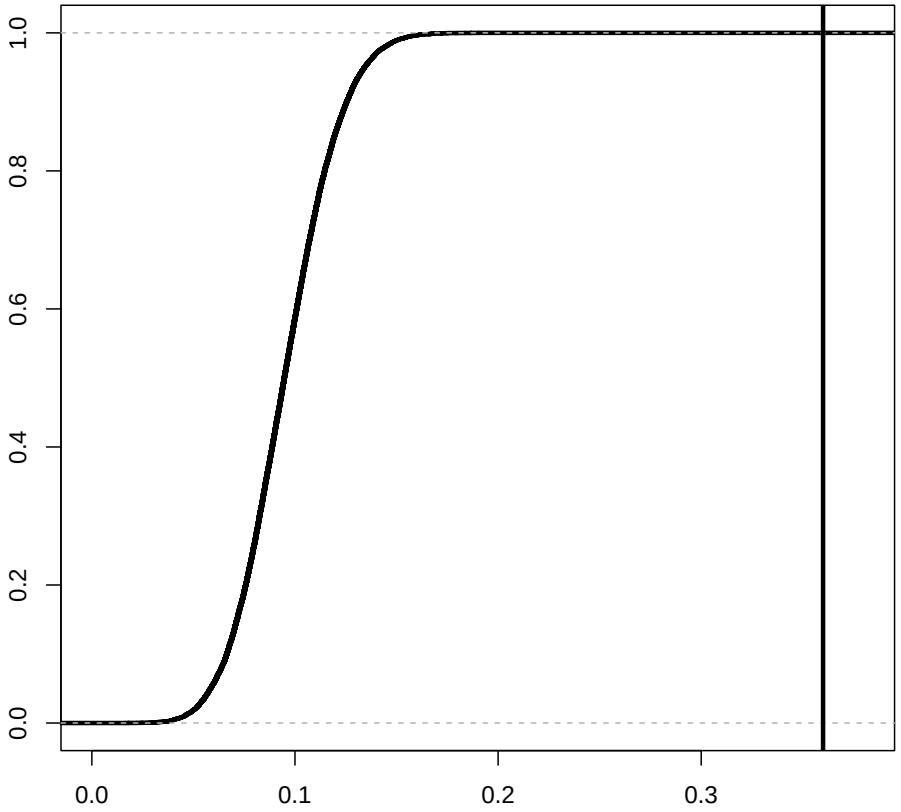


Figure 2: Left: Estimated number N_k of upsets with k extreme points (i.e., minimal elements, logarithmic display). Right: Distribution of the estimates of $|\mathcal{S}|$ (k is drawn uniformly from $\{1, \ldots, 33\}$ and then $|\mathcal{S}_k|$ is estimated and multiplied with 33. The expectation of the obtained estimator is then $|\mathcal{S}|$.

Comparison with permutation test

- permutation test with 20000 resamples (took ca. 12 min time)
- estimated p-value non-parametrically: 0 (or $\frac{1}{20000} = 5 \cdot 10^{-5}$)
- parametric p-value $\approx 4.8 \cdot 10^{-31}$



A last example

- $N = 200$ data points in $\mathbb{R}^{10}$.
- interordinaly scaled, i.e., we look at the closure system $\mathcal{S}$ of all 10-dimensional hypercubes generated by these 200 data points.
- naive analysis gives $|\mathcal{S}| \leq \left(N^{10}\right)^2 \approx 10^{46}$
- VC analysis gives: $\hbar \leq 2 \cdot 10 = 20$ and therefore $|\mathcal{S}| \leq 1.5 \cdot \frac{N^{\hbar}}{\hbar!} \approx 6.5 \cdot 10^{27}$
- concretely estimating $|\mathcal{S}|$ (for standard-normally distributed data points) gives $|\mathcal{S}| \approx 5 \cdot 10^{20}$

Reminder: Formal concept analysis

Reminder: Formal concept analysis (FCA)

Given: formal context $\mathbb{K} := (G, M, I)$ where

- G is a set of objects,
- M is a set of attributes,
- $I \subseteq G \times M$ is a binary relation with the interpretation $(g, m) \in I$ iff object g has attribute m.
- Aim: Describe I with the help of so-called formal concepts.
- Note: In the sequel, we will always assume that G is finite.

Definition (formal concept)

Let $\mathbb{K} := (G, M, I)$ be a formal context. Define for $A \subseteq G$ and $B \subseteq M$ the associated sets

$$A' = \{m \in M \mid \forall g \in A : (g, m) \in I\}$$
$$B' = \{g \in G \mid \forall m \in B : (g, m) \in I\}.$$

Then, a pair (A, B) where $A \subseteq G$ is a set of objects and $B \subseteq M$ is a set of attributes is called a formal concept if $B = A'$ and $A = B'$. In such a case, we call A the extent and B the intent of the formal concept (A, B).

Remarks I

- In the sequel, we will look at the family of all extents $A \subseteq G$. This family is a closure system (i.e., a family of sets that contains G and that is closed under arbitrary intersections). We will denote this closure system with $\mathcal{S}$.
- The operator that maps an arbitrary set $A \subseteq G$ to its so-called closure A'' (or hull, think of the convex hull of points in $\mathbb{R}^d$,) is denoted with γ.

Remarks II

- We will also use so-called formal (object) implications denoted by $A \longrightarrow B$ where $A, B \subseteq G$. A formal implication $A \longrightarrow B$ is valid in a formal context/closure system, if every extent/set of the closure system C that does contain all elements of the so-called premise A does also contain every element of the so-called conclusion B.
- Note that the **context** $\mathbb{K}$, the corresponding **closure system** $\mathcal{S}$ of concept extents, the corresponding **closure operator** γ and the family of all valid formal **object implications** are equivalent descriptions of the underlying data structure. Therefore, in the sequel, we will assign certain properties always interchangeably to all of these equivalent descriptions.

Remarks III

- In the sequel we will always assume that the context has no duplicates, i.e. objects $g, h \in G$ with identical attributes. (This is no hard restriction because one can always handle duplicates with a corresponding weighting of objects.)

Illustrations

Caution!

Not everything that is valid and intuitive in $\mathbb{R}^d$ does also hold for general contexts/closure systems/closure operators!

Minimal Generators

Definition (minimal generator)

A set $A \subseteq G$ is called a **minimal generator** if we have $\gamma(B) \subsetneq \gamma(A)$ for all $B \subsetneq A$. The set of all minimal generators is denoted with $mingen(\mathcal{S})$. The set of all minimal generators of size K is denoted with $mingen_K(\mathcal{S})$

Remark

A set A is a minimal generator iff for all $B \subseteq A$ with $|A \backslash B| = 1$ we have $\gamma(B) \subsetneq \gamma(A)$.

Theorem

The following statements are equivalent:

i) *A is a* **minimal generator**.

ii) *A is* **implcation free**, *i.e., there exists no valid implication $B \longrightarrow C$ with $B \subseteq A$, $\emptyset \neq C \subseteq A$ and $C \cap B = \emptyset$.*

iii) *A is* **shatterable** *w.r.t. $\mathcal{S} := \gamma(2^G)$, i.e. $A \cap \mathcal{S} := \{A \cap B \mid B \in \mathcal{S}\} = 2^A$.*

Definition (VC dimension)

The **VC dimension** of a closure system is defined as the maximal cardinality of a shatterable set (or a minimal generator or an implication free set).

Meet-distributive closure systems

Definition (meet-distributive closure system)

Let $\mathbb{K} := (G, M, I)$ be a formal context (without duplicates), let $\mathcal{S}$ be the corresponding closure system of all concept extents and let γ be the corresponding closure operator. Then $\mathbb{K}$ (or $\mathcal{S}$ or γ) is called **meet-distributive** if one of the following equivalent properties hold:

i) every closure $A \in \gamma(2^G)$ is the closure of its extreme points[a]

ii) the so-called anti-exchange property holds: For every $A \in \gamma(2^G)$ and $x, y \notin A$ we have

$$A \cup \{x\} \to \{y\} \quad \Longrightarrow \quad A \cup \{y\} \not\to \{x\}$$

[a] A point $x \in A$ is an extreme point of A if $A \backslash \{x\} \not\to \{x\}$.

Illustrations

Theorem

A finite context without duplicates is meet-distributive if and only if every closure $A \in \gamma(2^G)$ has exactly one minimal generator $B \subseteq A$ (i.e., $\gamma(B) = A$ and $\gamma(C) \subsetneq A$ for every $C \subsetneq B$), namely the set of all extreme points of A.

Counting concepts

Counting concepts

Simple idea: count minimal generators instead of concepts.
Because for meet-distributive contexts the mapping

$$\gamma^* : \mathit{mingen}(\mathcal{S}) \to \mathcal{S} : A \mapsto \gamma(A)$$

is a bijection, this will work. (The inverse mapping is given by

$$\mathit{extr} : \mathcal{S} \to \mathit{mingen}(S) : A \mapsto \mathit{extr}(A),$$

where $\mathit{extr}(A) := \{x \in A \mid x \text{ is extreme point of } A\}$.)

How to count minimal generators?

Definition (independence)

We say that a point $x \in G$ is **independent** of a set $A \subseteq G$ if $x \notin A$ and the set $A \cup \{x\}$ is implication free. For a fixed set A we denote with $ind(A)$ the set of all $x \in G$ that are independent of A.

Enumerating all minimal generators of size K:

Algorithm 1 Enumerating $mingen_K$

1: $result \leftarrow \emptyset$
2: **for** $g_1 \in G$ **do**
3: **for** $g_2 \in ind(\{g_1\})$ **do**
4: **for** $g_3 \in ind(\{g_1, g_2\})$ **do**
5: $\vdots$
6: **for** $g_K \in ind(\{g_1, g_2, \ldots, g_{K-1}\})$ **do**
7: $result \leftarrow result \cup \{g_1, \ldots, g_K\}$
8: **end for**
9: $\vdots$
10: **end for**
11: **end for**
12: **end for**
13: **return** *result*

Counting all minimal generators of size K:

Algorithm 2 Counting $mingen_K$

1: $size \leftarrow 0$
2: **for** $g_1 \in G$ **do**
3: **for** $g_2 \in ind(\{g_1\})$ **do**
4: **for** $g_3 \in ind(\{g_1, g_2\})$ **do**
5: $\vdots$
6: **for** $g_K \in ind(\{g_1, g_2, \ldots, g_{K-1}\})$ **do**
7: $size \leftarrow size + 1$
8: **end for**
9: $\vdots$
10: **end for**
11: **end for**
12: **end for**
13: **return** $\frac{size}{K!}$

Estimating the size of minimal generators of size K:

Algorithm 3 Counting $mingen_K$

1: $size \leftarrow 0$
2: **for** $g_1 \in G$ **do**
3: **for** $g_2 \in ind(\{g_1\})$ **do**
4: **for** $g_3 \in ind(\{g_1, g_2\})$ **do**
5: $\vdots$
6: **for** $g_K \in ind(\{g_1, g_2, \ldots, g_{K-1}\})$ **do**
7: $size \leftarrow size + 1$
8: **end for**
9: $\vdots$
10: **end for**
11: **end for**
12: **end for**
13: **return** $\frac{size}{K!}$

Estimating the size of minimal generators of size K:

Idea:

- In the steps 6-8 one computes the size of $ind(\{g_1, g_2, \ldots, g_{K-1}\})$
- In the steps 5-9 one computes the size of $ind(\{g_1, g_2, \ldots, g_{K-2})$ (up to double-counting)
- $\vdots$
- Now replace in every loop the exact counting scheme by a mean-unbiased estimator:

Algorithm 4 estimating $mingen_K$

1: $size \leftarrow 1$
2: **for** $g_1 \in G$ **do**
3: compute exactly $|ind(\{g_1\})|$
4: $size \leftarrow size \cdot |ind(\{g_1\})|$
5: **for** g_2 randomly drawn from $ind(\{g_1\})$ **do**
6: compute exactly $|ind(\{g_1, g_2\})|$
7: $size \leftarrow size \cdot |ind(\{g_1, g_2\})|$
8: $\vdots$
9: **for** g_{K-1} randomly drawn from $ind(\{g_1, g_2, \ldots, g_{K-2}\})$ **do**
10: compute exactly $|ind(\{g_1, g_2, \ldots, g_{K-1}\})|$
11: $size \leftarrow size \cdot |ind(\{g_1, g_2, \ldots, g_{K-1}\})|$
12: **end for**
13: $\vdots$
14: **end for**
15: **end for**
16: **return** $\frac{size}{K!}$ as a mean-unbiased estimator of $|mingen_K|$