

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

---

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

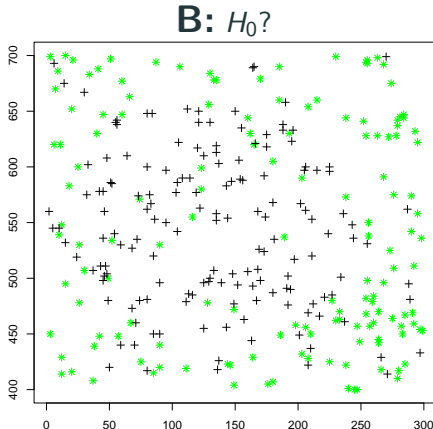
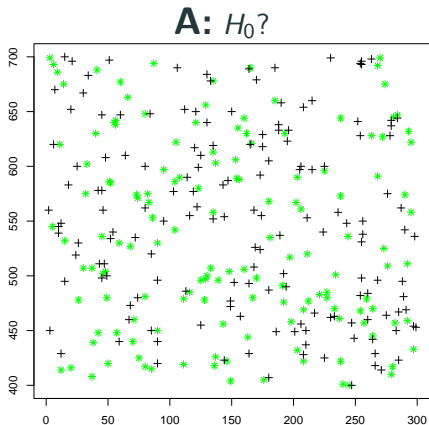
**1** A motivating example & some comments (on statistical learning theory/statistics)

**2** Further examples

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

---

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



$H_0$ : 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

$\mathcal{S} := \{A \cap X_{obs} \mid A \subseteq \mathbb{R}^2, A \text{ convex}\}$  (a **finite!** 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  $N := n_1 + n_2$ ;  $m := \max\{n_1, n_2\}$ ;

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

## Motivation/Question of interest

- ▶  $H_0 : P_1 = P_2$  vs.  $H_1 : P_1 \neq P_2$
- ▶ 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  $|S|$ .**  
(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)

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

$\underbrace{\hspace{15em}}_{=: \alpha(\varepsilon)}$

with  $N := n_1 + n_2$  (and assuming  $n_1 = n_2$ ) and  $\mathfrak{h}$  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  $n = n_1 = n_2$  :

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

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

Serfling:  $P(P(A) - \hat{P}^n(A) > \varepsilon) \leq \exp[-2n\varepsilon^2/(1 - f_n^*)]$   
(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  $c := \sup_{k \in \{1, \dots, L\}} P(A_k)$

## “Reminder” II

Triangle inequality (here):

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

Reason:

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

## Comments

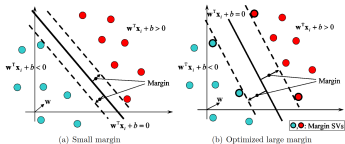
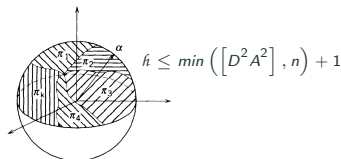
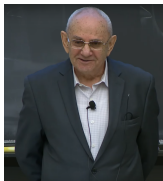


*“The reality is that the VC line of analysis leads to a very loose bound...”*

*“... 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



*“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]*



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

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

with  $\sigma_1, \dots, \sigma_N$  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  $S := \bigcup_{j \in I} S_j$  (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)



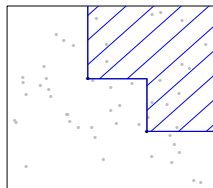
## Further examples

---

# Further examples of meet-distributive closure systems

## upsets

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

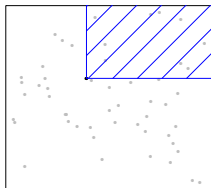


Application example:  
multivariate poverty-/  
inequality analysis (3  
dimensions income,  
education and health)

VC dimension = maximal  
number of corners

## principal filters

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

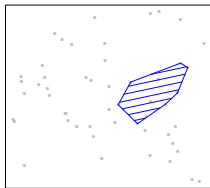


possible applications:  
multidimensional  
K.-S.-tests, e.g. item  
response theory (item  
impact or DIF)

VC dimension = 1  
FALSCH!!

## convex sets

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



possible applications:  
spatial statistics

VC dimension =  $\infty$  (or  
maximal number of  
extreme points)

## Further examples of meet-distributive closure systems

- ▶ Subgroup discovery with only interordinally scaled variables without ties.
- ▶ Apposition<sup>1</sup> 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)

---

<sup>1</sup>In the sense of combining many meet-distributive contexts to one context by using all attributes from every context.

## Example: convex sets

$$N = 300$$

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

$$\text{t-Cl: } [0; 1.9 \cdot 10^{16}]$$

$$\text{abc-Cl: } [10^{15}; 1.7 \cdot 10^{17}]$$

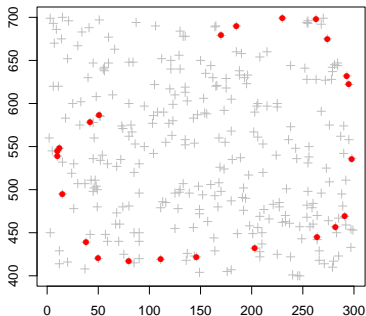
confidence level:  $1 - 10^{-10}$

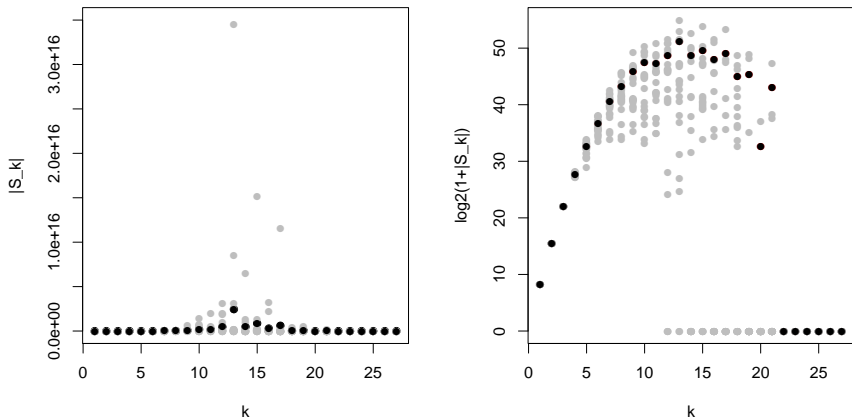
based on sample of size 540

computation time ca  $6h$

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

$$C_{\mathcal{V}} \geq 3.2 \cdot 10^{39}$$





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

## Example: convex sets: (some) $H_1$

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

$D_n^{obs} \approx 0.728$

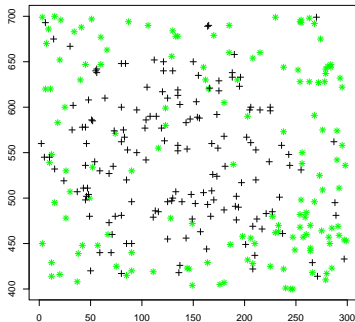
$P_{H_0}(D_n > D_n^{obs}) \leq \Gamma \approx 1.2 \cdot 10^{-15}$

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

runtime GUROBI (with tricks): 129s

$\hat{h} \geq 27$  (out of memory)

VC bound useless (too loose)



## Example: convex sets: $H_0$

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

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

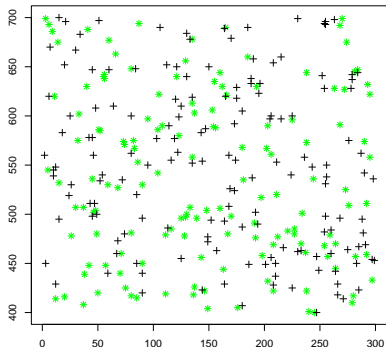
$$P_{H_0}(D_n > D_n^{obs}) \leq 5.2 \cdot 10^{14} ?$$

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

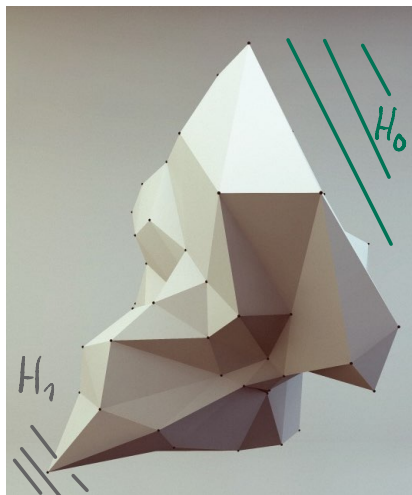
(out of memory)

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

for one resample

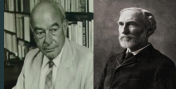


## Side remark





## 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<sup>2</sup>, Hoeffding's inequality (?) gives

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

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

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

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

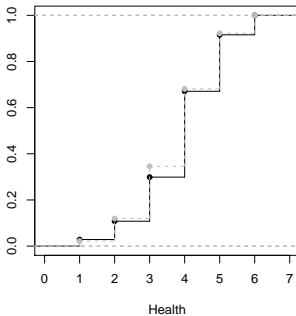
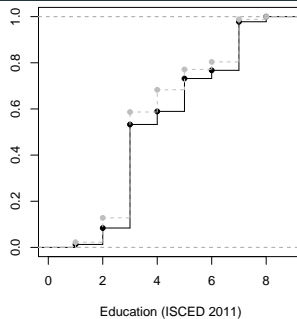
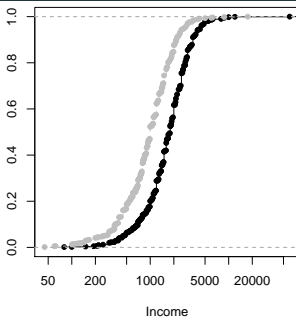
---

<sup>2</sup>A random vector  $(X_1, \dots, X_n)$  is called  $(r - 1)$  dependent if for  $j - i \geq r$  the random vectors  $(X_1, \dots, X_i)$  and  $(X_j, \dots, X_N)$  are independent.

## Another application example (upsets)

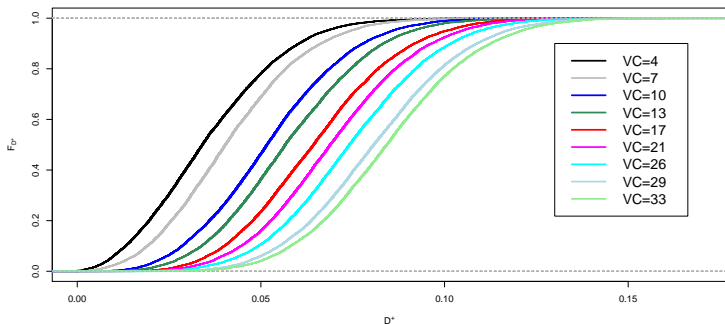
- ▶ Subsample of Allbus 2014 (706 female and 809 male respondents).
- ▶ Dimensions:
  - *Income*.
  - *Health* (self-reported, ranging from 1 (bad) to 6 (excellent)).
  - *Education* (ISCED 2011: ranging from 0 (less than primary education) to 8 (doctoral or equivalent level)).

# Marginal analysis



## Joint analysis

- ▶  $h = 33$  (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.)

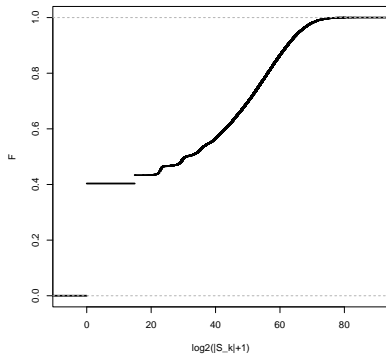
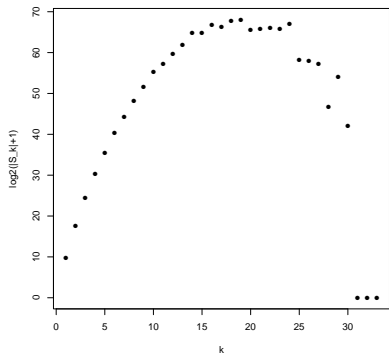


- ▶  $h = 33$ , therefore we have  $|A| \leq 33$  for every minimal generator  $A \in \text{mingen}(\mathcal{S})$  (see later).
- ▶ estimate of 3300 sampled minimal generators (took ca. 1.4 min time):

$$|\mathcal{S}| \approx 7 \cdot 10^{20}$$

$$CI : [0; 1.9 \cdot 10^{21}] \text{ (confidence level: } 1 - 10^{-10}\text{)}$$

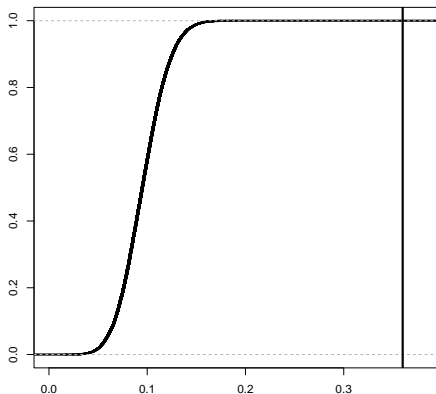
$$\Gamma \approx 1.9 \cdot 10^{-16} \quad (CI : [0; 5.2 \cdot 10^{-16}])$$



**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  $\{1, \dots, 33\}$  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  $\mathbb{R}^{10}$ .
- ▶ interordinally 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  $|\mathcal{S}| \leq (N^{10})^2 \approx 10^{46}$
- ▶ VC analysis gives:  $\hat{h} \leq 2 \cdot 10 = 20$  and therefore  $|\mathcal{S}| \leq 1.5 \cdot \frac{N^{\hat{h}}}{\hat{h}!} \approx 6.5 \cdot 10^{27}$
- ▶ concretely estimating  $|\mathcal{S}|$  (for standard-normally distributed data points) gives  $|\mathcal{S}| \approx 5 \cdot 10^{20}$