Fallibilistic regularization for (semi-inductive) abduction by adjoining auxiliary co-outcomes:

Rethinking Vapnik's 'Rethinking statistical learning theory: learning using statistical invariants' [Vapnik and Izmailov, 2019] in the context of social sciences and subgroup discovery

Georg Schollmeyer

18.03.2021

Oder kurz:

Annahmen vs. Hinzunahmen

(assumptions vs. adjunctions)

I: An example of subgroup discovery

III: Abduction & exploratory data analysis [Yu, 1994]

IV: Abduction and ill-posedness or the poverty of (automated) exploratory data analysis?

V: Regularization

I: An example of subgroup discovery

Determinants of political interest

Determinants according to Niedermayer [2005]:

- education
- occupation
- income
- sex
- age

Further determinants according to Hoecker [2013]:

- subjective social class membership

Further covariates:

- role model
- trust in the Bundestag
- immigrant

Response variable of interest: political interest

Subgroup discovery

Definition (subgroup discovery)

“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]

Subgroup discovery in the language of Formal concept analysis

Problem statement

Given a formal context $\mathbb{K} = (G, M, I)$, a target variable $y : G \longrightarrow \mathcal{A}$ and a quality function $q : 2^G \longrightarrow \mathbb{R}$, which "measures" the interestingness of subgroups of objects, the task of subgroup discovery is to find the best k formal concepts (A, B), for which the quality-values

$$q(A)$$

are as high as possible. (In the following: $k = 1$.)

Subgroup Discovery in our Situation

	x			y
	income $\leq$ 1000	sex=male	$\cdots$	poltical interest
person 1	×	×		0
person 2		×		1
person 3				0
$\vdots$				1
person m		×		0

Find subgroup A described by attributes (e.g., 'age between 20 and 45', sex ='female') with the statistically most interesting distribution of the target variable 'political interest'.

Mathematically i

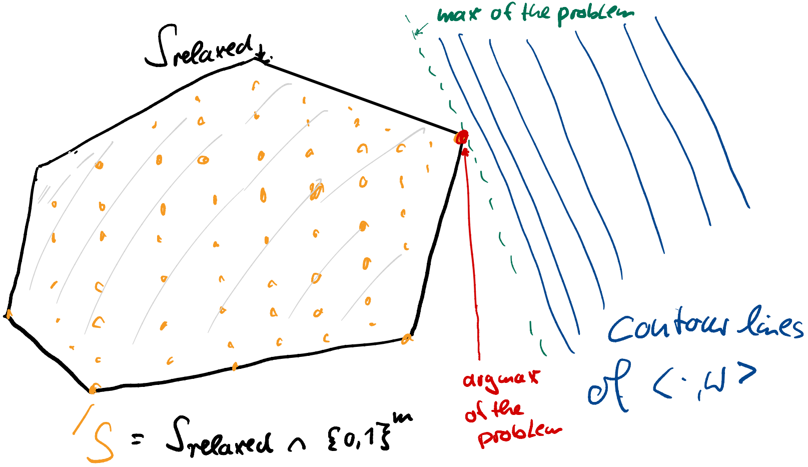
- $\mathcal{S}$ . . . family of all subgroups, i.e., family of all sets of persons that belong to a certain attribute description.
- More concisely: $\mathcal{S}$ . . . closure system of all concept extents of the given context.
- Let a vector $s \in \{0,1\}^m$ describe a subgroup A via $s_i = 1 \iff$ person i is in subgroup A.
- Then, the subgroup discovery problem can be stated as:

$$\langle s, w \rangle \longrightarrow \max$$

under $s \in \{0,1\}^m$ s.t. s is an indicator vector for some $A \in \mathcal{S}$

Mathematically ii

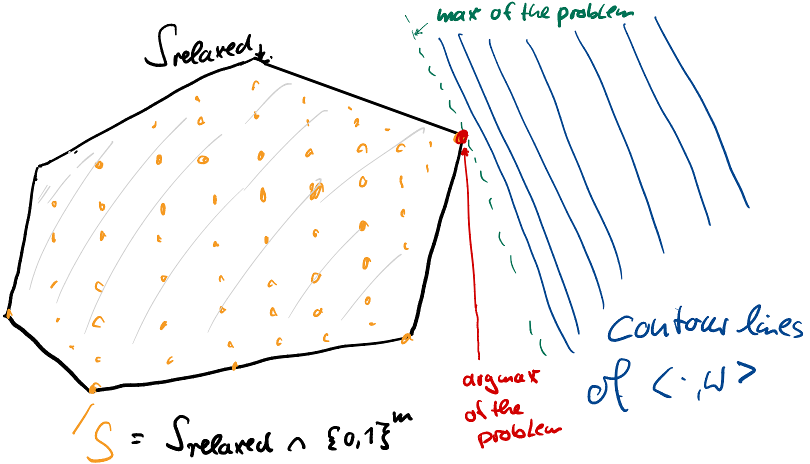
- with $w_i = \begin{cases} \frac{1}{\sum\limits_{i=1}^{m} y_i} & \text{if } y_i = 1 \\ \frac{-1}{m - \sum\limits_{i=1}^{m} y_i} & \text{if } y_i = 0 \end{cases}$.
- Here, w is a fixed vector which only depends on the target variable 'political interest'.
- The constraints for s (beyond $s \in \{0, 1\}^m$) can be formulated as linear constraints.



$$\langle s, w\rangle \longrightarrow \max$$

$$\text{under } s \in \{0,1\}$$

$$As \leq b$$

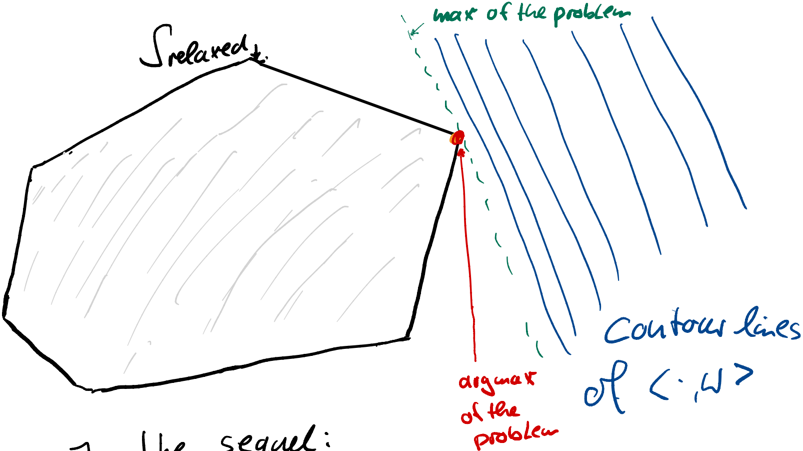


In terms of a statistical test:

$$T := \sup_{A \in \mathcal{S}} \langle s_A, w \rangle; \;\; T^* := \sup_{A \in \mathcal{S}} \langle s_A, w^* \rangle \quad (w^* \ldots \text{ population version of } w)$$

$H_0 : T^* = 0$ vs.

$H_1 : T^* > 0$



In the sequel:

ignore condition

$s \in \{0,1\}^m$ for

the sake of simplicity

of the illustrations

Problem

Often $\mathcal{S}$ very large. $\rightsquigarrow$ $T := \sup_{A \in \mathcal{S}} \langle s_A, w \rangle$ very ill-behaved.

(Note: In the following analysis we treat the covariates x as fixed and the target variable y as random.)

III: Abduction & exploratory data analysis [Yu, 1994]

Abduction

If μ were true, then π, π', π'' would follow as miscellaneous consequences;

$\underline{\text{But } \pi, \pi', \pi'' \text{ are in fact true;}}$

Provisionally, we may suppose that μ is true.

'*In short, for Peirce abductive reasoning is inference to an explanatory hypothesis, and this form of reasoning is to be distinguished from induction, in which we have already adopted a hypothesis and are only testing its consequences.*'[Campos, 2011, p.425]

Abduction

'*Abduction, the logic suggested by Peirce, can be viewed as a logic of exploratory data analysis. For Peirce abduction is the firstness (existence, actuality); deduction, the secondness (possibility, potentiality); and induction, the thirdness (generality, continuity). Abduction plays the role of generating new ideas or hypotheses; deduction functions as evaluating the hypotheses; and induction is justifying of the hypothesis with empirical data.*'[Yu, 1994]

IV: Abduction and ill-posedness or the poverty of (automated) exploratory data analysis?

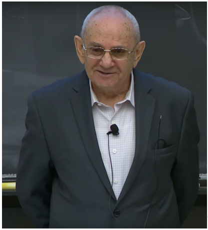
Abduction and ill-posedness or the poverty of (automated) exploratory data analysis?

'*In exploratory data analysis, after observing some surprising facts, we exploit them and check the predicted values against the observed values and residuals. Although there may be more than one convincing patterns, we 'abduct' only those which are more plausible. In other words,* **exploratory data analysis is not trying out everything.** *Rescher (1978) interpreted abduction as an opposition to Popper's falsification (1963). There are millions of possible explanations to a phenomenon. Due to the economy of research, we cannot afford to falsify every possibility. As mentioned before, we don't have to know everything to know something. By the same token,* **we don't have to screen every false thing to dig out the authentic one. Peirce argued that animals have the instinct to do the right things without struggling, we humans, as a kind of animal, also have the innate ability to make the right decision intuitively.**'

'*For Peirce, progress in science depends on the observation of the right facts by minds furnished with appropriate* **ideas** *(Tursman, 1987)...*'

'*...In short, abduction by intuition, can be interpreted as observing the world with appropriate* **categories** *which arise from the internal structure of meanings. The implications of abduction for researchers is that the use of exploratory data analysis is neither exhausting all possibilities nor making hasty decisions. Researchers must be well equipped with proper categories in order to sort out the* **invariant features** *and patterns of phenomena. The statistical method, in this sense, is not only number crunching, but also a thoughtful way of dissecting data.*' [Yu, 1994]

The poverty of (automated) exploratory data analysis?



[a] *'The next three decades (1970s, 1980s, and 1990s) were crucial for developments in statistics. After the shocking discovery that the classical approach suffers from the curse of dimensionality, statisticians tried to find methods that could replace classical methods in solving real-life problems. During this time statistics was split into two very different parts: theoretical statistics that continued to develop the classical paradigm of generative models, and applied statistics that suggested a compromise between theoretical justification of the algorithms and heuristic approaches to solving real-life problems.*

[a] Vladimir Vapnik: Complete Statistical Theory of Learning: MIT Deep Learning Series. January 2020. `https://www.youtube.com/watch?v=Ow25mjFjSmg`

The poverty of (automated) exploratory data analysis?

*They tried to justify such a position by inventing special names for such activities (***exploratory data analysis***), where in fact the superiority of common sense over theoretical justification was declared. However, they never tried to construct or justify new algorithms using VC theory. Only after SVM technology became a dominant force in data mining methods did they start to use its technical ideas (but not its philosophy) to modify classical algorithms.*[1]' [Vapnik, 2006]

[1] Statisticians did not recognise conceptual aspects of VC theory. Their criticism of this theory before SVM was that the VC bounds were too loose to be useful. Therefore the theory is not practical and to create new methods it is better to use common sense than the results of this theory.

The poverty of (automated) exploratory data analysis?

1) Testing an abduced hypothesis with the same data that were used for the abductive step is clearly inductively invalid (in the first place). (trivial)

The poverty of (automated) exploratory data analysis?

2) Testing an abduced hypothesis with new data is inductively valid but most probably useless in an 'overfitting' / 'ill-posed' situation.
$\leadsto$ Regularization has to be implemented in the abductive step.
(trivial, too)

The poverty of (automated) exploratory data analysis?

The actual poverty

3) But how to wisely regularize if one does not know beforehand what one will abductively find in the data? (non-trivial, if solvable at all.)

(Remember: '*... we don't have to screen every false thing to dig out the authentic one. Peirce argued that animals have the instinct to do the right things without struggling, we humans, as a kind of animal, also have the innate ability to make the right decision intuitively.*' [Yu, 1994] What about automata?)

V: Regularization

- Canonical (blind?) regularization
- Regularization qua general principles (continuity, smoothness)
- Bayesian regularization
- The regularization from here:

Data-driven vs. data-overdriven Regularization

A too simple remedy:

- Idea: reduce $\mathcal{S}$ with the help of the actually observed data.
- for example, take $\mathcal{S}^* \subsetneq \mathcal{S}$, with $\mathcal{S}^*$ consisting of that subgroups that look 'appropriate' in the light of the data.
- extreme case : $\mathcal{S}^* = \mathcal{S}^*(y) := \{A \mid \langle s_A, w\rangle \geq c \cdot T\}$ (e.g., with $c = 1$)
- Obvious: $\sup\limits_{A \in \mathcal{S}^*} \langle s_A, w\rangle = \sup\limits_{A \in \mathcal{S}} \langle s_A, w\rangle$.
- $\rightsquigarrow$ useless!

Other resolutions

a) Gather more data

b) Regularize canonically

c) Use further **assumptions**: Utilize additional 'justified' / 'empirically corroborated' theory/knowledge and classical statistical modelling together with auxiliary constructs measured by auxiliary data

d) Use **adjunctions**: Regularize with the help of auxiliary co-outcomes

a) More data

Vertically more data

covariates	political interest
xxxx....x............xx	0
x.......xx.........xx......	1
xxx.......xx.xxxxx	0
xxx..................	1
....x........xxx......xx.	1

⇝

covariates	political interest
xxxx....x............xx	0
x.......xx.........xx......	1
xxx.......xx.xxxxx	0
xxx..................	1
....x........xxx......xx.	1
xxx...xxxxx..x.x.xxx	1
xxx.x.x.xxxxx.x.	0
xxx.x..........	0
.....xx.x.x.x.x.x......	1
.....x.....x..	0
xx...x.x.x..x.x..x..	1
xx.....x..x..	1
......xxx..x..x	0

Note: $\mathcal{S}$ gets larger.

a) More data

Horizontally more data

covariates	political interest
xxxx....x............xx	0
x.......xx.........xx......	1
xxx.......xx.xxxxx	0
xxx..................	1
....x........xxx......xx.	1

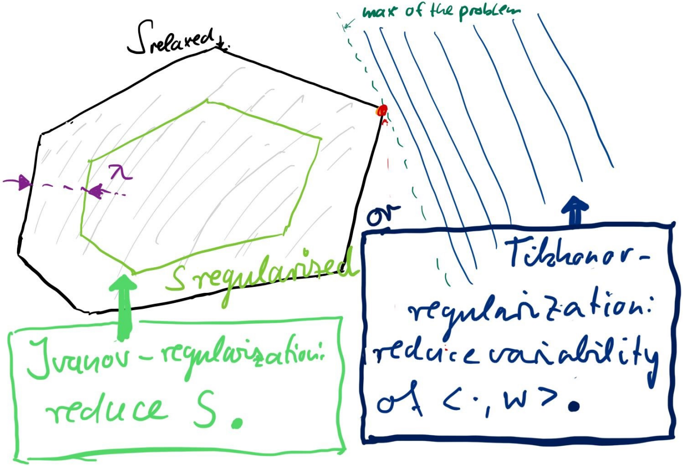
$\rightsquigarrow$

covariates	political interest	z^1 ?	...?	z^p?
xxxx....x............xx	0			
x.......xx.........xx......	1			
xxx.......xx.xxxxx	0			
xxx..................	1			
....x........xxx......xx.	1			

- Note: $\mathcal{S}$ does not get larger.
- But which auxiliary variables to use (repeated measurements?) and how?
- In fact, for the way we proceed in resolution d) the family $\mathcal{S}$ will get smaller.

b) Canonical Regularization

- How to regularize in a very discrete setting?
- Not straight-forward (but possible)
- Most easy thing to do: 'Canonical' regularization



$S_{relaxed}$
max of the problem
λ
$S_{regularised}$
or
Ivanov-regularization: reduce S.
Tikhonov-regularization: reduce variability of $\langle\cdot, w\rangle$.

c) Statistical modeling assuming auxiliary constructs: Rasch-Tikhonov type regularization

- Assume that there is a latent construct C which is measured by the variable political interest y.
- Additionally, assume that there are further variables $z^1, \ldots, z^p$ that do also measure the latent trait C.
- Assume a concrete latent-trait model, e.g. the Rasch model for the variables $y, z^1, \ldots, z^p$.
- Then given these assumptions are true, it would be wise to use all variables $y, z^1, \ldots, z^p$ to measure C and then to do a subgoup discovery with C as a target variable instead of y. (Concretely, one would replace y by the estimated solution-probability obtained through the estimated construct C.)

c) Statistical modeling assuming auxiliary constructs: Rasch-Tikhonov type regularization

Reminder: Rasch model:

$$\{0,1\} \ni X_{kI} = 1 \iff \text{person } k \text{ solves item } i.$$

$$P(X_{kI} = 1) = logis(\Theta_k - \sigma_i) \quad \text{with}$$

$$logis(a) = \frac{exp(a)}{1 + \exp(a)} \quad \text{and}$$

$$\Theta_k \ldots \text{ ability parameter of person } k$$

$$\sigma_i \ldots \text{ difficulty parameter of item } i \quad \text{and}$$

$$X_{ki} \perp X_{lj} \quad \text{for } (k,i) \neq (l,j)$$

c) Statistical modeling assuming auxiliary constructs: Rasch-Tikhonov type regularization

Mathematically: replace

$$T = \sup_{A \in \mathcal{S}} \langle s_A, w \rangle \quad \text{by} \quad T^{reg} = \sup_{A \in \mathcal{S}} \langle s_A, w^{reg} \rangle$$

with w^{reg} according to $y_i^{reg} := logis(\hat{\theta}(y_i + z_i^1 + \cdots + z_i^p))$ and $\hat{\theta} : \mathbb{R} \longrightarrow \mathbb{R}$ the isotone ML-estimator of the person ability parameter and $logis \circ \hat{\theta}$ approximately a positive affine map. (Note: In the Rasch model the row sums are sufficient statistics for the ability parameters. Note further that for identical dificulty parameters $logis \circ \hat{\theta} = id$.) More 'generally', one can take

$$T^{reg} = \sup_{A \in \mathcal{A}} \langle s_A, w \rangle + \sum_{i=1}^{p} \lambda_i \cdot \langle s_A, w_{z^i}^{reg} \rangle,$$

thus the name Rasch-Tikhonov type 'regularization'.

c) Statistical modeling assuming auxiliary constructs: Rasch-Tikhonov type regularization

- **Assume** that there is a latent construct C which is measured by the variable political interest y.
- Additionally, **assume** that there are further variables $z^1, \ldots, z^p$ that do also measure the latent trait C.
- **Assume** a concrete latent-trait model,e.g. the Rasch model for the variables $y, z^1, \ldots, z^p$.
- Then given these **assumptions** are true, it would be wise to use all variables $y, z^1, \ldots, z^p$ to measure C and then to do a subgoup discovery with C as a target variable instead of y. (Concretely, one would replace y by the estimated solution-probability obtained through the estimated construct C.)

However, assuming that all assumptions are fulfilled may be very naive:

Example

- Allbus 2018
- $z^1, \ldots, z^5$:

  '*Hier haben wir noch ein kurzes Quiz mit Fragen zur Politik. Manche Fragen sind eher einfach, andere eher schwierig.*'

  Zu welcher Partei gehören die folgenden Politiker und Politikerinnen?

  Antwortmöglichkeiten:

  -8 Weiß nicht

  1 CDU bzw. CSU

  2 Die Grünen

  3 Die Linke

  4 FDP

  5 SPD

  6 AfD

Example

- Allbus 2018
- $z^1, \ldots, z^5$:

  '*Hier haben wir noch ein kurzes Quiz mit Fragen zur Politik. Manche Fragen sind eher einfach, andere eher schwierig.*'

  Zu welcher Partei gehören die folgenden Politiker und Politikerinnen?

  Politiker:

  z^1: Heiko Maas

  z^2: Christian Lindner

  z^3: Peter Altmaier

  z^4: Katrin Göring-Eckhardt

  z^5: Angela Merkel

  $\vdots$

d) Structural risk minimization by adjoining auxiliary co-outcomes: Vapnik-Ivanov type regularization

Idea:

- Do not assume anything more but adjoin potentially useful co-outcomes $z^1, \dots, z^p$ that are in relation to y in the sense that subgroups where y has e.g., a high proportion of 1's but the z^i's have a low or medium proportion of 1's are not as 'interesting' (in the second place) as subgroups with a high proportion of 1's both for y, as well as for the z^i's.
- Note: Because of the ill-posedness of the unregularized problem we necessarily have to reduce $\mathcal{S}$. We use the additional variables $z^1, \dots, z^p$ only to guide the process of reducing $\mathcal{S}$.
- Note: We are still interested in y and not in the z^i's (and also not in any (contrived) latent construct).

d) Structural risk minimization by adjoining auxiliary co-outcomes: Vapnik-Ivanov type regularization

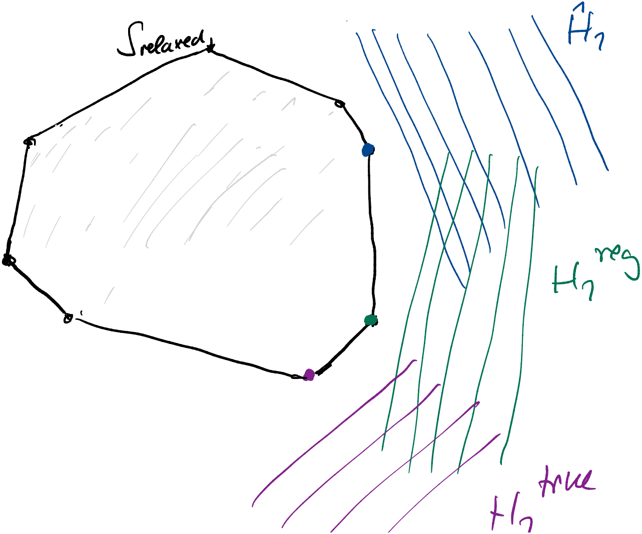
Mathematically: replace

$$T = \sup_{A \in \mathcal{S}} \langle s_A, w \rangle \quad \text{by}$$
$$T = \sup_{A \in \mathcal{S}^{reg}} \langle s_A, w \rangle$$

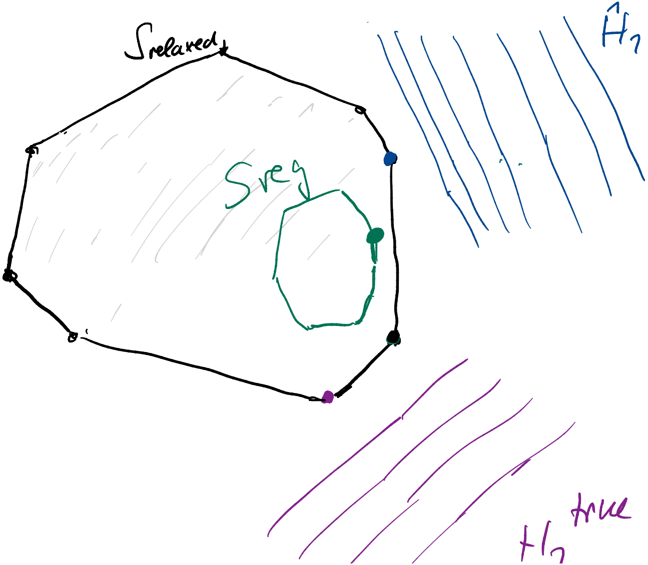
with $\mathcal{S}^{reg} := \{A \in \mathcal{S} \mid \langle s_A, w_{z^1}^{reg} \rangle \geq c^1, \ldots, \langle s_A, w_{z^p}^{reg} \rangle \geq c^p\} \subseteq \mathcal{S}$.

Thus the name Vapnik-Ivanov type regularization.

Rasch-Thikonov vs. Vapnik-Ivanov regularization



Rasch-Thikonov vs. Vapnik-Ivanov regularization



General Duality of Tikhonov and Ivanov type regularization

folklore: For $L, p : D \longrightarrow \mathbb{R}$, $\lambda > 0$ under certain uniqueness-assumptions, the following statements are equivalent:

(1) x maximizes $L + \lambda p$ over D

(2) x maximizes L under the constraint

$p(x) \geq C^*$ with

$C^* = p(x^*)$ and

x^* the maximizer of (1).

Note: In our context, the constant C^* that mediates the duality is random.

General Duality of Tikhonov and Ivanov type regularization

What does this mean for the interrelations between resolutions c) and d)?

Inductive validity: Inference

- permutation test?
  What is the appropriate Null hypothesis?
  - $X \perp Y$ (not enough)
  - $X \perp (Y, Z^1, \dots, Z^p)$
  - $(X, Z^1, \dots, Z^p) \perp Y$
  - $Y \mid X = x, Z^1 = z^1, \dots, Z^p = z^p$ does not depend on x
    or shortly: $Y \perp X \mid (Z^1, \dots, Z^p)$
- sample splitting as an ultimate solution?
- random covariate case?

Literatur

D. G. Campos. On the distinction between peirce's abduction and lipton's inference to the best explanation. *Synthese*, 180(3): 419–442, 2011.

B. Hoecker. *Frauen und das institutionelle Europa: politische Partizipation und Repräsentation im Geschlechtervergleich*. Springer-Verlag, 2013.

O. Niedermayer. *Politische Orientierungen*, pages 16–155. VS Verlag für Sozialwissenschaften, Wiesbaden, 2005. ISBN 978-3-322-80815-8. doi: 10.1007/978-3-322-80815-8_2. URL `https://doi.org/10.1007/978-3-322-80815-8_2`.

V. Vapnik and R. Izmailov. Rethinking statistical learning theory: learning using statistical invariants. *Machine Learning*, 108(3): 381–423, 2019.

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.

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

C. H. Yu. Abduction? deduction? induction? Is There a Logic of Exploratory Data Analysis?. 1994. ERIC. Institute of Education Sciences.