The Statistics-“Machine” in Data Science

Peter Bühlmann
ETH Zürich

Acknowledgments

very special honor

main collaborators:

Nicolai Meinshausen
ETH Zurich

> Es ist wahnsinnig spannend und macht Spass.

Sara van de Geer
ETH Zurich

fathers:

Hansruedi Künsch
my doctoral father

Hans Bühlmann
my “true” father

my wife and my family

Data Science

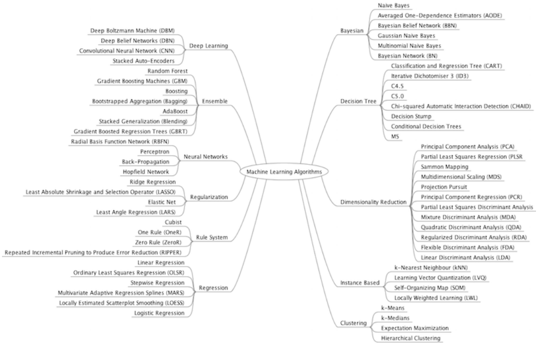
lots and lots of data

how certain are we that conclusions inferred from data “hold”?

often heard nowadays:

“... and we then apply (interpretable) machine learning” to

- predict
- classify
- gain understanding of the system
- infer the causes



Machine Learning Algorithms

Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders

Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)

Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network

Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)

Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)

Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5

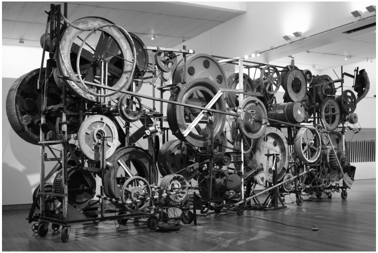
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)

Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)

Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

$\leadsto$ it’s a collection of tools/methods/algorithms!

Why not a Statistics-“Machine”?’



a collection of tools and methods
for inferential and “confirmatory” statements

perhaps it’s nothing new:
except the issue of dealing with more complex data

and perhaps a bit a marketing slogan that
“statistics” is also a key player in Data Science

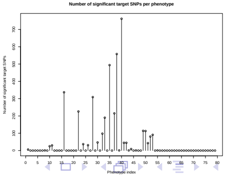
An example: Behavioral economics and genetics

joint project with Ernst Fehr, Univ. Zurich

- $n = 1'525$ persons
- genetic information (SNPs): $p \approx 10^6$
- 79 response variables, measuring “behavior”

$$p \gg n$$

goal: find significant associations between behavioral responses and genetic markers



... and let's have a look at *Nature 496, 398 (25 April 2013)*

Challenges in irreproducible research

...

"the complexity of the system and of the techniques ... do not stand the test of further studies"

- "We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data."
- "We will also demand more precise descriptions of statistics, and we will commission statisticians as consultants on certain papers, at the editors discretion and at the referees suggestion."
- "Too few budding scientists receive adequate training in statistics and other quantitative aspects of their subject."

... and let's have a look at *Nature 496, 398 (25 April 2013)*

Challenges in irreproducible research

...

"the complexity of the system and of the techniques ... do not stand the test of further studies"

- "We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data."
- "We will also demand more precise descriptions of statistics, and we will commission statisticians as consultants on certain papers, at the editors discretion and at the referees suggestion."
- "Too few budding scientists receive adequate training in statistics and other quantitative aspects of their subject."

what we aim for:

- assessment of uncertainty, replicability and generalizability
- meaningful statements towards “causality”

  does the value of a biomarker “causally influence” e.g. risk aversion?

regarding the example on “behavioral economics and genetics”:

inferential statements are difficult, due to the very high-dimensional nature of the problem

not yet “Big Data” (only a million variables, thousands of sample points)

in fact, so far: GWAS (genome-wide association study) are usually based on marginal correlations between a response and genetic variables

only correlation $\rightsquigarrow$ can be very spurious!

regarding the example on “behavioral economics and genetics”:

inferential statements are difficult, due to the
very high-dimensional nature of the problem

not yet “Big Data” (only a million variables, thousands of sample points)

in fact, so far: GWAS (genome-wide association study) are usually based on marginal correlations between a response and genetic variables
only correlation ⇝ can be very spurious!

only correlation $\rightsquigarrow$ can be very spurious! (Messerli, 2012)

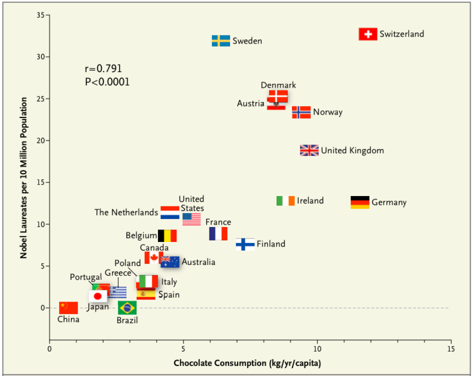


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

(empirical)
correlation = 0.791 !

X: chocolate consumption per capita (per yr.)
Y: number of Nobel Laureates per 10 million popul.

Linear model: the statistical workhorse for getting beyond correlations

$$\underbrace{Y_i}_{\text{response } i\text{th obs.}} = \sum_{j=1}^{p} \beta_j^0 \underbrace{X_i^{(j)}}_{j\text{th covariate } i\text{th. obs.}} + \underbrace{\varepsilon_i}_{i\text{th error term}} \quad , i = 1, \ldots, n$$

standard vector- and matrix-notation:

$$\begin{aligned} & Y_{n\times 1} = X_{n\times p}\beta^0_{p\times 1} + \varepsilon_{n\times 1} \\ \text{in short :} \quad & Y = X\beta^0 + \varepsilon \end{aligned}$$

- design matrix X: either deterministic or stochastic
- error/noise ε:
  $\varepsilon_1, \ldots, \varepsilon_n$ independent, $\mathbb{E}[\varepsilon_i] = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma_i^2 \leq \sigma^2$
  ε_i uncorrelated from X_i (when X is stochastic)

interpretation:

β_j^0 measures the effect of $X^{(j)}$ on Y when
“conditioning on” the other covariables $\{X^{(k)};\ k \neq j\}$

that is: it measures the effect of $X^{(j)}$ on Y which is not explained by the other covariables
$\rightsquigarrow$ much more a “causal” interpretation

equivalent to partial correlation and
very different from (marginal) correlation between $X^{(j)}$ and Y

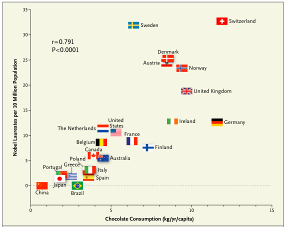


Figure 1. Correlation between Countries’ Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Regularized parameter estimation

$$Y = X\beta^0 + \varepsilon, \quad p \gg n$$

ℓ_1-norm regularization

(Tibshirani, 1996; Chen, Donoho and Saunders, 1998)

also called Lasso (Tibshirani, 1996):

$$\hat{\beta}(\lambda) = \text{argmin}_\beta(n^{-1}\|Y - X\beta\|_2^2 + \lambda \underbrace{\|\beta\|_1}_{\sum_{j=1}^p |\beta_j|})$$

convex optimization problem

- sparse solution (because of "ℓ_1-geometry")
  that is: many $\hat{\beta}_j(\lambda) = 0$
- not unique in general... but unique with high probability under some assumptions (which we make "anyway")

LASSO = Least Absolute Shrinkage and Selection Operator

Near-optimal statistical properties of Lasso (for fixed design X)

assumptions:

- identifiability:
  note $X\beta^0 = X\theta$ for any $\theta = \beta^0 + \xi$, ξ in the null-space of X
  $\rightsquigarrow$ restricted eigenvalue or compatibility condition
  van de Geer (2007); Bickel, Ritov & Tsybakov (2009); van de Geer & PB (2009);...
  weaker than RIP (Candes & Tao, 2006)
- sparsity: let $S_0 = \mathrm{supp}(\beta^0) = \{j;\ \beta_j^0 \neq 0\}$ and assume
  $s_0 = |S_0| = o(n/\log(p))$ (or $o(\sqrt{n/\log(p)})$)
- sub-Gaussian error distribution

$\rightsquigarrow$ with high probability, and choosing $\lambda \asymp \sqrt{\log(p)/n}$

$$\|\hat{\beta} - \beta^0\|_2^2 = O(s_0 \log(p)/n),\ \|\hat{\beta} - \beta^0\|_1 = O(s_0\sqrt{\log(p)/n}),$$
$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O(s_0 \log(p)/n)$$

(PB & van de Geer (2011), Hastie, Tibshirani & Wainwright (2015),...)

$\rightsquigarrow$ Lasso: a most popular method in high-dimensional statistics

Near-optimal statistical properties of Lasso (for fixed design X)

assumptions:

- identifiability:
  note $X\beta^0 = X\theta$ for any $\theta = \beta^0 + \xi$, ξ in the null-space of X
  $\leadsto$ restricted eigenvalue or compatibility condition
  van de Geer (2007); Bickel, Ritov & Tsybakov (2009); van de Geer & PB (2009);...
  weaker than RIP (Candes & Tao, 2006)
- sparsity: let $S_0 = \mathrm{supp}(\beta^0) = \{j;\ \beta^0_j \neq 0\}$ and assume
  $s_0 = |S_0| = o(n/\log(p))$ (or $o(\sqrt{n/\log(p)})$)
- sub-Gaussian error distribution

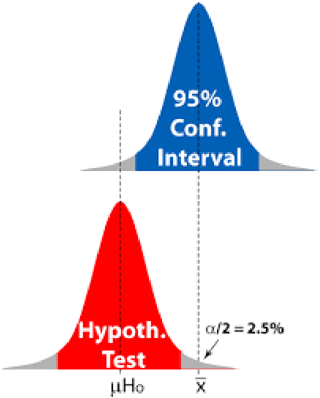
$\leadsto$ with high probability, and choosing $\lambda \asymp \sqrt{\log(p)/n}$

$$\|\hat{\beta} - \beta^0\|_2^2 = O(s_0 \log(p)/n),\ \|\hat{\beta} - \beta^0\|_1 = O(s_0\sqrt{\log(p)/n}),$$
$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O(s_0 \log(p)/n)$$

(PB & van de Geer (2011), Hastie, Tibshirani & Wainwright (2015),...)

$\leadsto$ Lasso: a most popular method in high-dimensional statistics

Uncertainty quantification: p-values and confidence intervals



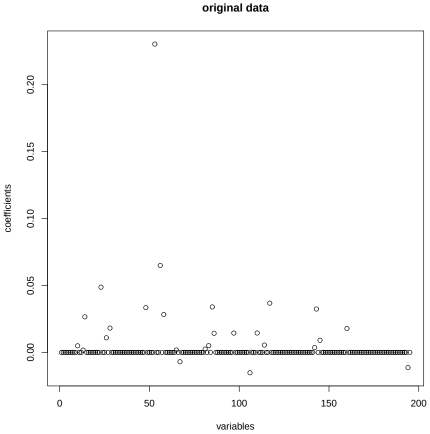
frequentist
uncertainty quantification

(in contrast to Bayesian inference)

- use classical concepts but in high-dimensional non-classical settings
- develop less classical things $\rightsquigarrow$ hierarchical inference
- ...

Toy example: Motif regression ($p = 195, n = 143$)

Lasso estimated coefficients $\hat{\beta}(\hat{\lambda}_{\mathrm{CV}})$



p-values/quantifying uncertainty would be very useful!

$$Y = X\beta^0 + \varepsilon \;\; (p \gg n)$$

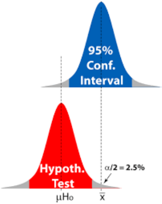
classical goal: statistical hypothesis testing

$$H_{0,j} : \beta_j^0 = 0 \text{ versus } H_{A,j} : \beta_j^0 \neq 0$$

or $$H_{0,G} : \beta_j^0 = 0 \;\forall\, j \in \underbrace{G}_{\subseteq\{1,\ldots,p\}} \text{ versus } H_{A,G} : \exists j \in G \text{ with } \beta_j^0 \neq 0$$

background: if we could handle the asymptotic distribution of the Lasso $\hat{\beta}(\lambda)$ under the null-hypothesis

$\leadsto$ could construct p-values



this is very difficult!
asymptotic distribution of $\hat{\beta}$ has some point mass at zero,...
Knight and Fu (2000) for $p < \infty$ and $n \to \infty$

because of "non-regularity" of sparse estimators
"point mass at zero" phenomenon $\rightsquigarrow$ "super-efficiency"

(Hodges, 1951)

$\rightsquigarrow$ standard bootstrapping and subsampling should not be used

Low-dimensional projections and bias correction (Zhang & Zhang, 2014)
Or de-sparsifying the Lasso estimator (van de Geer, PB, Ritov & Dezeure, 2014)

motivation (for $p < n$):

$\hat{\beta}_{\mathrm{LS},j}$ from projection of Y onto residuals $(X_j - X_{-j}\hat{\gamma}^{(j)}_{\mathrm{LS}})$

projection not well defined if $p > n$
$\leadsto$ use "regularized" residuals from Lasso on X-variables

$$Z_j = X_j - X_{-j}\hat{\gamma}^{(j)}_{\mathrm{Lasso}}$$

using $Y = X\beta^0 + \varepsilon \rightsquigarrow$

$$Z_j^T Y = Z_j^T X_j \beta_j^0 + \sum_{k \neq j} Z_j^T X_k \beta_k^0 + Z_j^T \varepsilon$$

and hence

$$\frac{Z_j^T Y}{Z_j^T X_j} = \beta_j^0 + \underbrace{\sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \beta_k^0}_{\text{bias}} + \underbrace{\frac{Z_j^T \varepsilon}{Z_j^T X_j}}_{\text{noise component}}$$

$\rightsquigarrow$ de-sparsified Lasso:

$$\hat{b}_j = \frac{Z_j^T Y}{Z_j^T X_j} - \underbrace{\sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \hat{\beta}_{\mathrm{Lasso};k}}_{\text{Lasso-estim. bias corr.}}$$

$\{\hat{b}_j\}_{j=1}^{p}$ is not sparse!... and this is crucial for Gaussian limit and it is "optimal" (see next)

- target: low-dimensional component β_j^0
- $\eta := \{\beta_k^0;\ k \neq j\}$ is a high-dimensional nuisance parameter
  $\rightsquigarrow$ exactly as in semiparametric modeling!
  and sparsely estimated (e.g. with Lasso)

Asymptotic pivot and optimality

Theorem (van de Geer, PB, Ritov & Dezeure, 2014)

$$\frac{\sqrt{n}(\hat{b}_j - \beta_j^0)}{\sigma_\varepsilon \sqrt{\Omega_{jj}}} \Rightarrow \mathcal{N}(0,1) \;\text{ as } p \geq n \to \infty$$

Ω_{jj} explicit expression $\sim (\Sigma^{-1})_{jj}$ optimal!

reaching semiparametric information bound

$\rightsquigarrow$ asympt. optimal p-values and confidence intervals
if we assume:

- population $\mathrm{Cov}(X) = \Sigma$ has minimal eigenvalue $\geq M > 0$ $\surd$
- sparsity for regr. Y vs. X: $s_0 = o(\sqrt{n}/\log(p))$"quite sparse"
- sparsity of design: Σ^{-1} sparse
  i.e. sparse regressions X_j vs. X_{-j}: $s_j \leq o(\sqrt{n/\log(p)})$
  may not be realistic
- no beta-min assumption !
  $\min_{j \in S_0} |\beta_j^0| \gg s_0 \sqrt{\log(p)/n}$ (or $s_0 \log(p)/n$)

Asymptotic pivot and optimality

Theorem (van de Geer, PB, Ritov & Dezeure, 2014)

$$\frac{\sqrt{n}(\hat{b}_j - \beta_j^0)}{\sigma_\varepsilon \sqrt{\Omega_{jj}}} \Rightarrow \mathcal{N}(0,1) \text{ as } p \geq n \to \infty$$

Ω_{jj} explicit expression $\sim (\Sigma^{-1})_{jj}$ optimal!

reaching semiparametric information bound

$\leadsto$ asympt. optimal p-values and confidence intervals
if we assume:

- population $\mathrm{Cov}(X) = \Sigma$ has minimal eigenvalue $\geq M > 0$ $\surd$
- sparsity for regr. Y vs. X: $s_0 = o(\sqrt{n}/\log(p))$"quite sparse"
- sparsity of design: Σ^{-1} sparse
  i.e. sparse regressions X_j vs. X_{-j}: $s_j \leq o(\sqrt{n/\log(p)})$
  may not be realistic
- no beta-min assumption !
  $\min_{j \in S_0} |\beta_j^0| \gg s_0\sqrt{\log(p)/n}$ (or $s_0 \log(p)/n$)

It is optimal!
Cramer-Rao

for data-sets with $p \approx 4'000 - 10'000$ and $n \approx 100$
$\leadsto$ often no significant variable

because
"β_j^0 is the effect when conditioning on all other variables..."

for example:
cannot distinguish between highly correlated variables $X^{(j)}, X^{(k)}$
but can find them as a significant group of variables where

> at least one among $\{\beta_j^0, \beta_k^0\}$ is $\neq 0$
> but unable to tell which of the two is different from zero

Behavioral economics and genomewide association

with Ernst Fehr, University of Zurich

- $n = 1525$ probands (all students!)
- $m = 79$ response variables measuring various behavioral characteristics (e.g. risk aversion) from well-designed experiments
- biomarkers: $\approx 10^6$ SNPs

model: multivariate linear model

$$\underbrace{\mathbf{Y}_{n\times m}}_{\text{responses}} = \underbrace{X_{n\times p}}_{\text{SNP data}} \beta^0_{p\times m} + \underbrace{\varepsilon_{n\times m}}_{\text{error}}$$

$$\mathbf{Y}_{n \times m} = X_{n \times p} \beta^0_{p \times m} + \varepsilon_{n \times m}$$

interested in p-values for

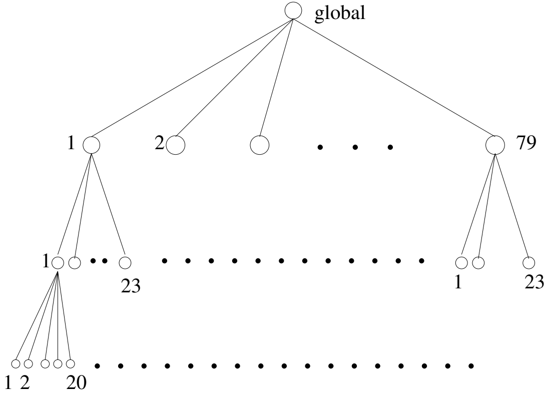
$$H_{0,jk}: \ \beta^0_{jk} = 0 \text{ versus } H_{A,jk}: \ \beta^0_{jk} \neq 0,$$
$$H_{0,G}: \ \beta^0_{jk} = 0 \text{ for all } j, k \in G \text{ versus } H_{A,G} = H^c_{0,G}$$

adjusted for multiple testing (among $\ell = O(10^6)$ hypotheses)

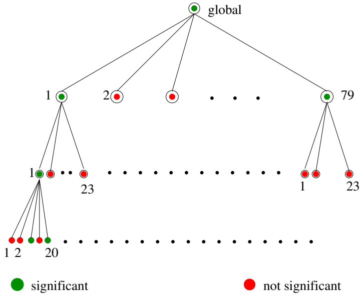
- standard: Bonferroni-Holm adjustment $\rightsquigarrow$ p-value $P_G \to P_{G;adj} = P_G \cdot \ell = P_G \cdot O(10^6)$!!!
- we want to do something much more efficient (statistically and computationally)

there is structure!

- 79 response experiments
- 23 chromosomes per response experiment
- groups of highly correlated SNPs per chromosome



do hierarchical FWER adjustment (Meinshausen, 2008)

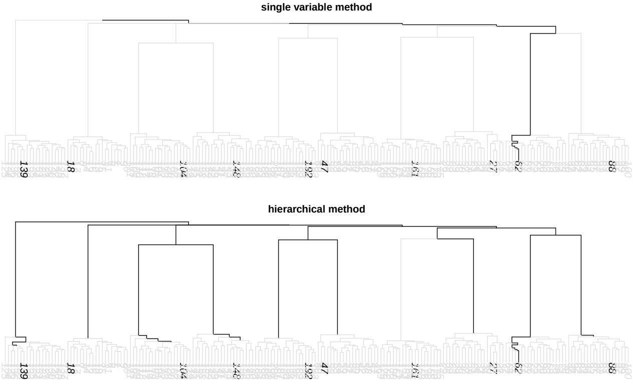


1. test global hypothesis
2. if significant: test all single response hypotheses
3. for the significant responses: test all single chromosome hyp.
4. for the significant chromosomes: test all groups of SNPs

$\rightsquigarrow$ powerful multiple testing with
data dependent adaptation of the resolution level

cf. general sequential testing principle (Goeman & Solari, 2010)

Mandozzi & PB (2015, 2016):



single variable method

hierarchical method

a hierarchical inference method is able to find additional groups of (highly correlated) variables

Sequential rejective testing: an old principle

(Marcus, Peritz & Gabriel, 1976)

ℓ hypothesis tests, ordered sequentially with hypotheses:

$$H_1 \prec H_2 \prec \ldots \prec H_\ell$$

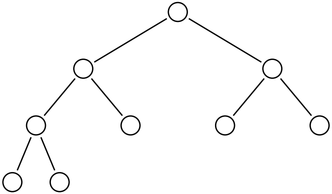
the rule:

- hypotheses are always tested on significance level α (no adjustment!)
- if H_r not rejected: stop considering further tests ($H_{r+1}, \ldots, H_\ell$ will not be considered)

easy to prove that

$$\text{FWER} = \mathbb{P}[\text{at least one false rejection}] \leq \alpha$$

in the context of hierarchical (e.g. binary) tree:



“essentially”:

- $H_1 \leftrightarrow$ top node of the tree $\leadsto$ level α
- $H_2 \leftrightarrow$ the 2 nodes of the second level of the tree
  $\leadsto$ do Bonferroni adjustment over 2 nodes
  $\leadsto$ level $\alpha/2$
- at a any level of depth in the tree: the sum of the levels $= \alpha$
  on each level of depth: Bonferroni correction

input:

- a hierarchy of groups/clusters $G \subseteq \{1, \ldots, p\}$
- valid p-values P_G for group testing
  use de-sparsified Lasso with test-statistics $\max_{j \in G} \frac{|\hat{b}_j|}{\widehat{s.e.}_j}$

$$H_{0,G}: \ \beta_j^0 = 0 \ \forall j \in G \ \text{ vs. } \ H_{A,G}: \ \beta_j^0 \neq 0 \text{ for some } j \in G$$

the essential operation is very simple:

$$P_{G;\mathrm{adj}} = P_G \cdot \frac{p}{|G|}, \quad P_G = \text{ p-value for } H_{0,G}$$

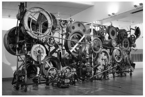
$$P_{G;\mathrm{hier-adj}} = \max_{D \in \mathcal{T}; G \subseteq D} P_{G;\mathrm{adj}} \quad \text{(“stop when not rejecting at a node”)}$$

$\leadsto$ the FWER is controlled (Meinshausen, 2008)

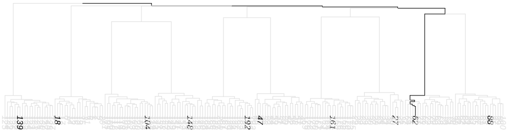
$$\mathbb{P}[\text{at least one false rejection}] \leq \alpha$$

the main benefit is not primarily the “efficient” multiple testing adjustment

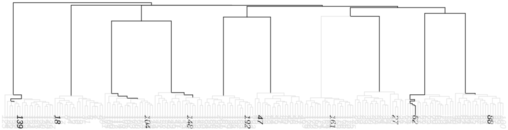
it is the fact that we automatically (data/machine-driven) adapt to an appropriate resolution level of the groups



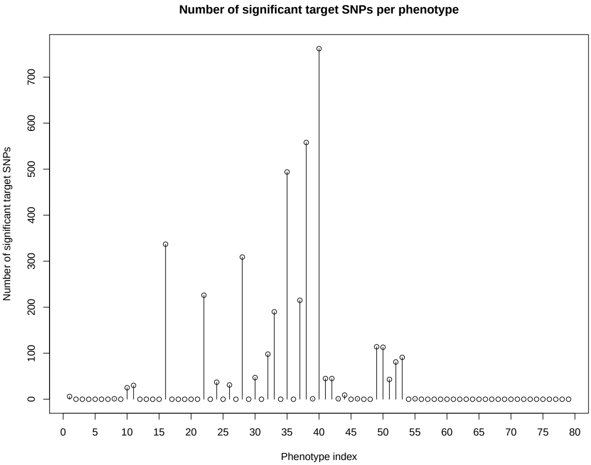
single variable method



hierarchical method



Behavioral economics example: number of significant SNP parameters per response



response 40 (?): most significant groups of SNPs

Genomewide association studies in medicine

a case for hierarchical inference!

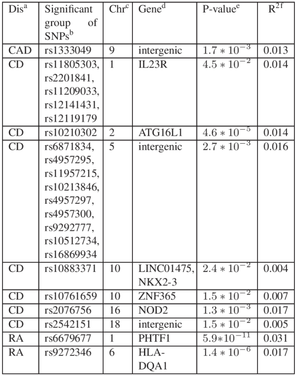
where the ground truth is much better known
(Buzdugan, Kalisch, Navarro, Schunk, Fehr & PB, 2016)

The Wellcome Trust Case Control Consortium (2007)

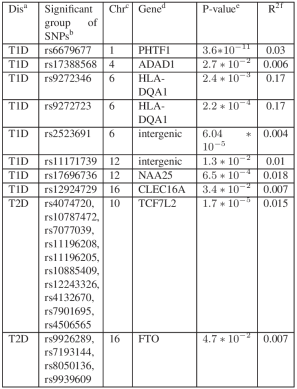
- 7 major diseases
- after missing data handling:
  2934 control cases
  about $1700 - 1800$ diseased cases (depend. on disease)
  approx. $p = 380'000$ SNPs per individual

coronary artery disease (CAD); Crohn's disease (CD); rheumatoid arthritis (RA); type 1 diabetes (T1D); type 2 diabetes (T2D)

significant small groups and single ! SNPs



Dis[a]	Significant group of SNPs[b]	Chr[c]	Gene[d]	P-value[e]	R^2[f]
CAD	rs1333049	9	intergenic	$1.7 * 10^{-3}$	0.013
CD	rs11805303, rs2201841, rs11209033, rs12141431, rs12119179	1	IL23R	$4.5 * 10^{-2}$	0.014
CD	rs10210302	2	ATG16L1	$4.6 * 10^{-5}$	0.014
CD	rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934	5	intergenic	$2.7 * 10^{-3}$	0.016
CD	rs10883371	10	LINC01475, NKX2-3	$2.4 * 10^{-2}$	0.004
CD	rs10761659	10	ZNF365	$1.5 * 10^{-2}$	0.007
CD	rs2076756	16	NOD2	$1.3 * 10^{-3}$	0.017
CD	rs2542151	18	intergenic	$1.5 * 10^{-2}$	0.005
RA	rs6679677	1	PHTF1	$5.9 * 10^{-11}$	0.031
RA	rs9272346	6	HLA-DQA1	$1.4 * 10^{-6}$	0.017



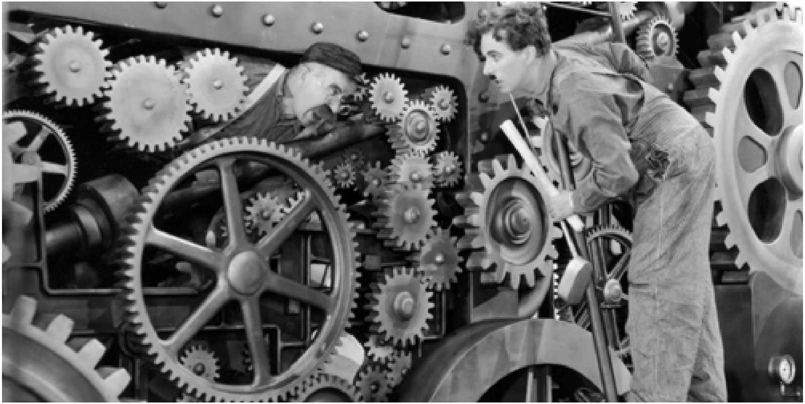
Dis[a]	Significant group of SNPs[b]	Chr[c]	Gene[d]	P-value[e]	R^2[f]
T1D	rs6679677	1	PHTF1	$3.6 * 10^{-11}$	0.03
T1D	rs17388568	4	ADAD1	$2.7 * 10^{-2}$	0.006
T1D	rs9272346	6	HLA-DQA1	$2.4 * 10^{-3}$	0.17
T1D	rs9272723	6	HLA-DQA1	$2.2 * 10^{-4}$	0.17
T1D	rs2523691	6	intergenic	$6.04 * 10^{-5}$	0.004
T1D	rs11171739	12	intergenic	$1.3 * 10^{-2}$	0.01
T1D	rs17696736	12	NAA25	$6.5 * 10^{-4}$	0.018
T1D	rs12924729	16	CLEC16A	$3.4 * 10^{-2}$	0.007
T2D	rs4074720, rs10787472, rs7077039, rs11196208, rs11196205, rs10885409, rs12243326, rs4132670, rs7901695, rs4506565	10	TCF7L2	$1.7 * 10^{-5}$	0.015
T2D	rs9926289, rs7193144, rs8050136, rs9939609	16	FTO	$4.7 * 10^{-2}$	0.007

for bipolar disorder (BD) and hypertension (HT): only large significant groups (containing between 1'000 - 20'000 SNPs) findings:

- recover some “established” associations:
  - single “established” SNPs
  - small groups containing an “established” SNP

  “established”: SNP is found by WTCCC or by WTCCC replication studies
- infer some significant non-reported groups
- automatically infer whether a disease exhibits high or low resolution associations to
  - high resolution: single or a small groups of SNPs (CAD, CD, RA, T1D, T2D)
  - low resolution: large groups of SNPs only (BD, HT)

Inspect the Statistics-“Machine”!



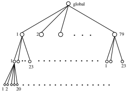
An experimental validation: Genomewide association study in plant biology

collaboration with Max Planck Institute for Plant Breeding Research (Köln):

Klasen, Barbez, Meier, Meinshausen, PB, Koornneef, Busch & Schneeberger (2016)

root development in Arabidopsis Thaliana
response Y: root size (root meristem zone-length)
$n = 201$, $p = 214'051$

hierarchical inference: 4 new significant small groups



3 new associations are within and neighboring to PEPR2 gene
$\leadsto$ validation: wild-type versus pepr2-1 loss-of-function mutant
which indeed resulted to impact root size (p-value 0.0007)
p-value = 0.0007 in Gaussian ANOVA model with 4 replicates

Towards Causality – which is a very ambitious word

we should think about (external) interventions
$\leadsto$ more mechanistsic and less “philosophical” approach

causality – an answer to a “what if I do” question:

if we would intervene on a gene, would this have an effect on a response of interest?

want to predict the outcome Y of such an intervention experiment withoug having data from such interventions

Towards Causality – which is a very ambitious word

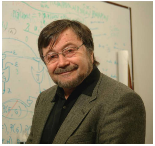
we should think about (external) interventions
$\rightsquigarrow$ more mechanistsic and less “philosophical” approach

causality – an answer to a “what if I do” question:

if we would intervene on a gene, would this have an effect on a response of interest?

want to predict the outcome Y of such an intervention experiment withoug having data from such interventions

... can be formalized with Pearl’s $\mathrm{do}(\cdot)$ operator



Judea Pearl, Turing Award 2011

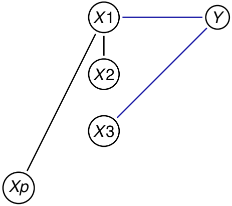
$$\begin{aligned} \text{Causal effect} &= \text{effect of an outside intervention/manipulation} \\ &= \text{effect seen in a randomized trial} \end{aligned}$$

we want to infer/predict causal effects from non-interventional (= observational) data? $\leadsto$ it's extrapolation!

for example in Policy making

James Heckman, Nobel Prize Economics 2000 technically:

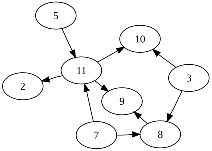
- regression effects are undirected associations



undirected edge $X^{(j)} - Y$ $\Leftrightarrow$ $\underbrace{\beta_j^0}_{j\text{th regr. coeff.}} \neq 0$

$\Leftrightarrow$ $Y, X^{(j)}$ conditionally dependent given all other $\{X^{(k)};\ k \neq j\}$

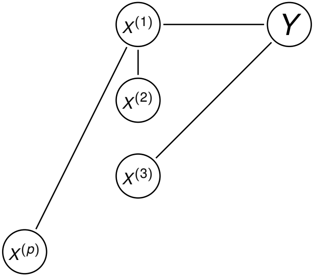
- causal effects are based on directed associations



directed edges describe
the causal influence diagram

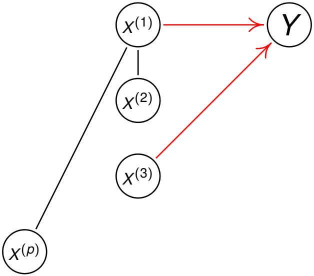
we simply postulate that effects (undirected edges) must point from genetic variables to disease status (and not vice-versa)

"everybody" would agree with this postulate



we simply postulate that effects (undirected edges) must point from genetic variables to disease status (and not vice-versa)

“everybody” would agree with this postulate



Proposition (nothing new at all)

Assume linear structural equation model where Y has no descendants (no children, no outgoing edges). Then:

$$X^{(j)} \rightarrow Y \;\Leftrightarrow\; \underbrace{\beta_j^0}_{j\text{th regr. coeff.}} \neq 0.$$

regression (almost) does the job!

regression (almost) does the job!

indeed: our significance in regression leads to an experimentally validated intervention effect

PEPR2 gene intervention leads to effect on root size
$\leadsto$ "causal" effect of PEPR2 gene

"almost": beware of hidden confounders...

but see Peters, PB & Meinshausen (2016)

regression (almost) does the job!

indeed: our significance in regression leads to an experimentally validated intervention effect

PEPR2 gene intervention leads to effect on root size
$\rightsquigarrow$ “causal” effect of PEPR2 gene

“almost”: beware of hidden confounders...

but see Peters, PB & Meinshausen (2016)

I am running out of time and cannot explain the details

Conclusions

The Statistics-“Machine” in Data Science:

- has deep historical roots, is very broad

many contributed!

- it enables uncertainty quantification, even in complex high-dimensional settings
- it contributes towards obtaining new scientific insights and “causal mechanisms”
- it benefits from other disciplines

in particular from Optimization and Comp. Sci.

THANK
YOU!

Crohn's disease

large groups

SNP group size	chrom.	p-value
3622	1	0.036
7571	2	0.003
18161	3	0.001
6948	4	0.028
16144	5	0.007
8077	6	0.005
12624	6	0.019
13899	7	0.027
15434	8	0.031
18238	9	0.003
4972	10	0.036
14419	11	0.013
11900	14	0.006
2965	19	0.037
9852	20	0.032
4879	21	0.009

most chromosomes exhibit signific. associations

no further resolution to finer groups

Toy example (Messerli, 2012): two variables

$$X = \text{ annual chocolate consumption per capita in a country}$$

$$Y = \text{ number of Nobel Prizes in a country}$$

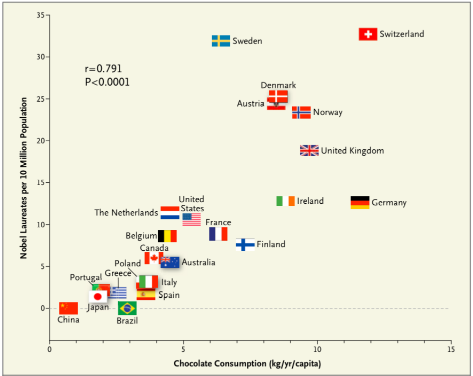


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

(empirical)
correlation = 0.791 !

Franz H. Messerli
Swiss cardiologist specializing in treatment of hypertension
honorary doctorate from Jagiellonian University Krakow (2013)

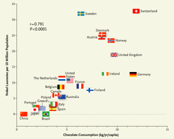


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- if we intervene on chocolate consumption

(and force everybody to eat the double amount of chocolate in Switzerland, on average: 24.7 → 49.4 grams per day...)

⇒ would the number of Nobel prizes go up?

- if we intervene on the number of Nobel prizes

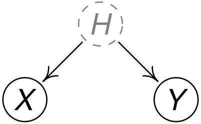
(hard to do – suppose we could manipulate award committee)

⇒ would the amount of chocolate consumption go up?

probably: both interventions would exhibit no effect

$\leadsto$ no "causal"/intervention relation between X and Y

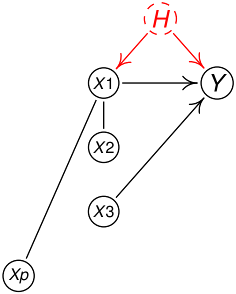
but there might be a hidden confounding variable H such as "social welfare/richness" which induces correlation



GWAS is a lucky situation! regression will (almost) do the job

except when:

- model is incorrect (e.g. interaction effects)
  can deal with model misspecification to a certain extent
  (PB & van de Geer, 2015)
- hidden confounder between SNPs and response

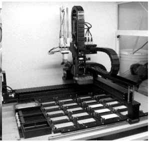


$\rightsquigarrow$ still an open problem in the context of GWAS

but see Peters, PB & Meinshausen (2016)

we also have gene deletion validations in yeast-biology

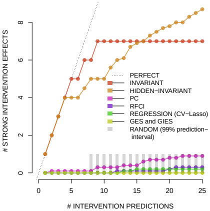
Meinshausen, Hauser, Mooij, Peters, Versteeg & PB, (2016)



ROC-type plot: “the steeper up the curve the better”

I : causal invariant prediction method

H: ... including hidden variables


# STRONG INTERVENTION EFFECTS
# INTERVENTION PREDICTIONS
PERFECT
INVARIANT
HIDDEN−INVARIANT
PC
RFCI
REGRESSION (CV−Lasso)
GES and GIES
RANDOM (99% prediction−interval)