Hierarchical High-Dimensional Statistical Inference

Peter Bühlmann
ETH Zürich

main collaborators:
Sara van de Geer, Nicolai Meinshausen, Lukas Meier, Ruben Dezeure, Jacopo Mandozzi, Laura Buzdugan

The motivation for this work: Games and Behavior, Economics and Genetics

James Francis Hannan

Hannan did very fundamental work in the theory of repeated games, compound decision problems, and mathemat. statistics

here, the human behavior w.r.t. games is measured
and one asks whether it is caused by genetics

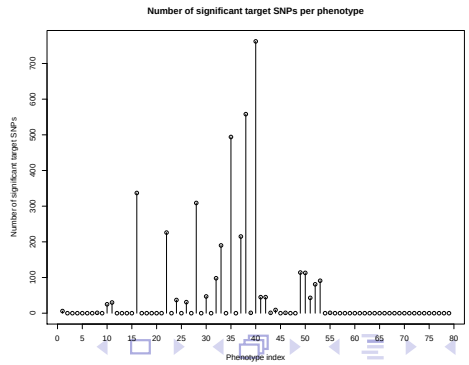
High-dimensional data

Behavioral economics and genetics (with Ernst Fehr, U. 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."

Linear model

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

standard vector- and matrix-notation:

$$Y_{n\times 1} = X_{n\times p}\beta^0_{p\times 1} + \varepsilon_{n\times 1}$$

$$\text{in short :} \qquad Y = X\beta^0 + \varepsilon$$

- 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
much more a "causal" interpretation

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

Regularized parameter estimation

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")
- 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

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
  (weaker than RIP)
- 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

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

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

$\rightsquigarrow$ Lasso is a standard workhorse in high-dimensional statistics

Near-optimal statistical properties of Lasso

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
  (weaker than RIP)
- 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

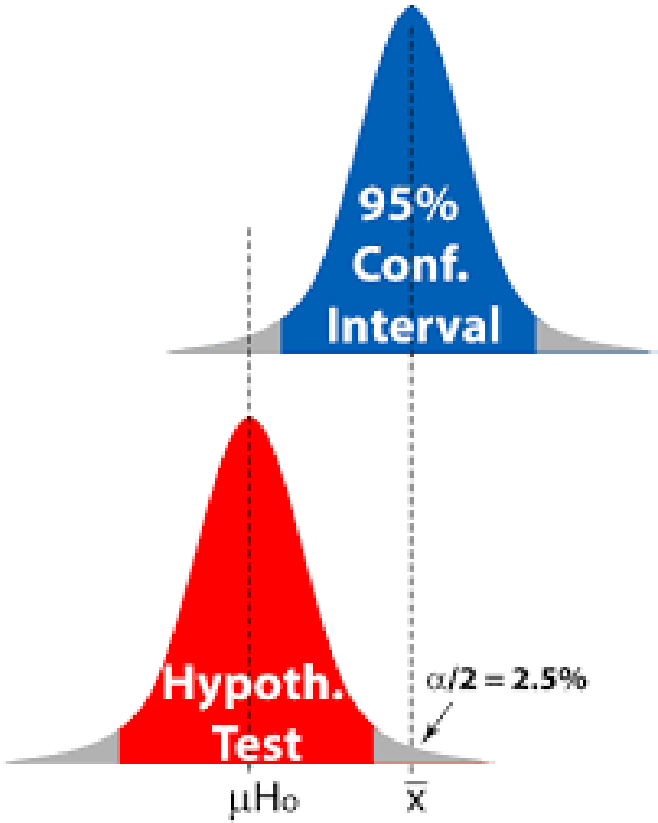
$\rightsquigarrow$ with high probability

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

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

$\rightsquigarrow$ Lasso is a standard workhorse in high-dimensional statistics

Uncertainty quantification: p-values and confidence intervals



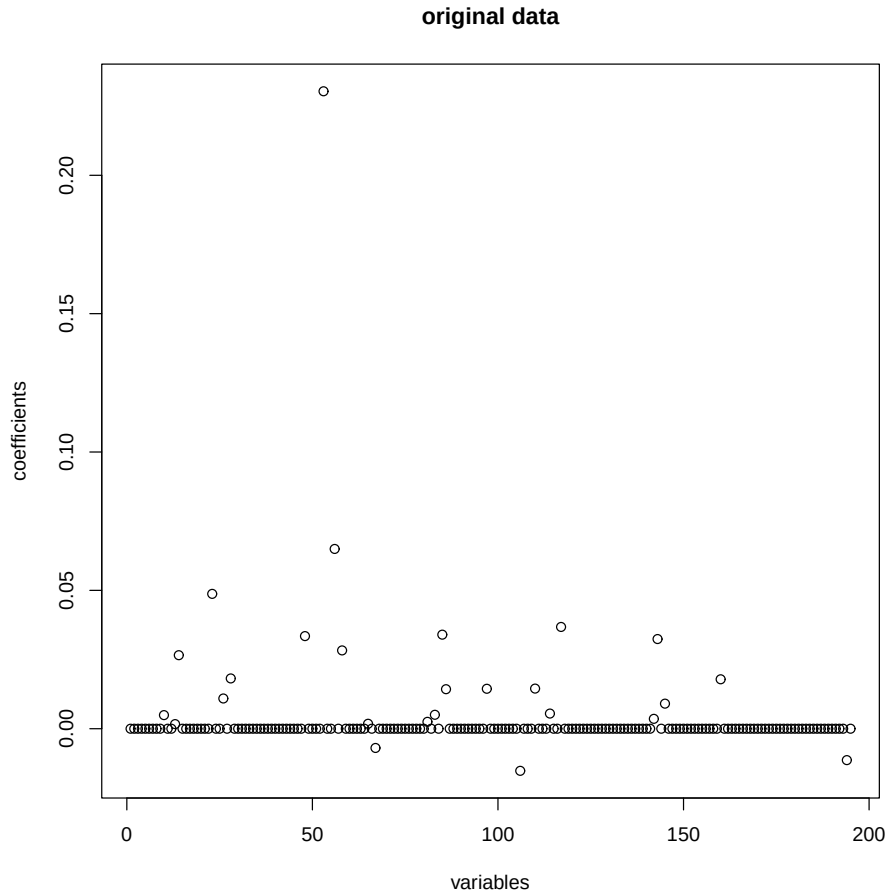
frequentist
uncertainty quantification

(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!

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

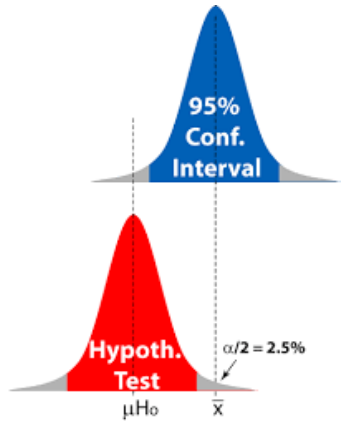
classical goal: statistical hypothesis testing

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

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

$\rightsquigarrow$ 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 $p > n$
$\rightsquigarrow$ 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:

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

$\{\hat{b}_j\}_{j=1}^p$ 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)

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

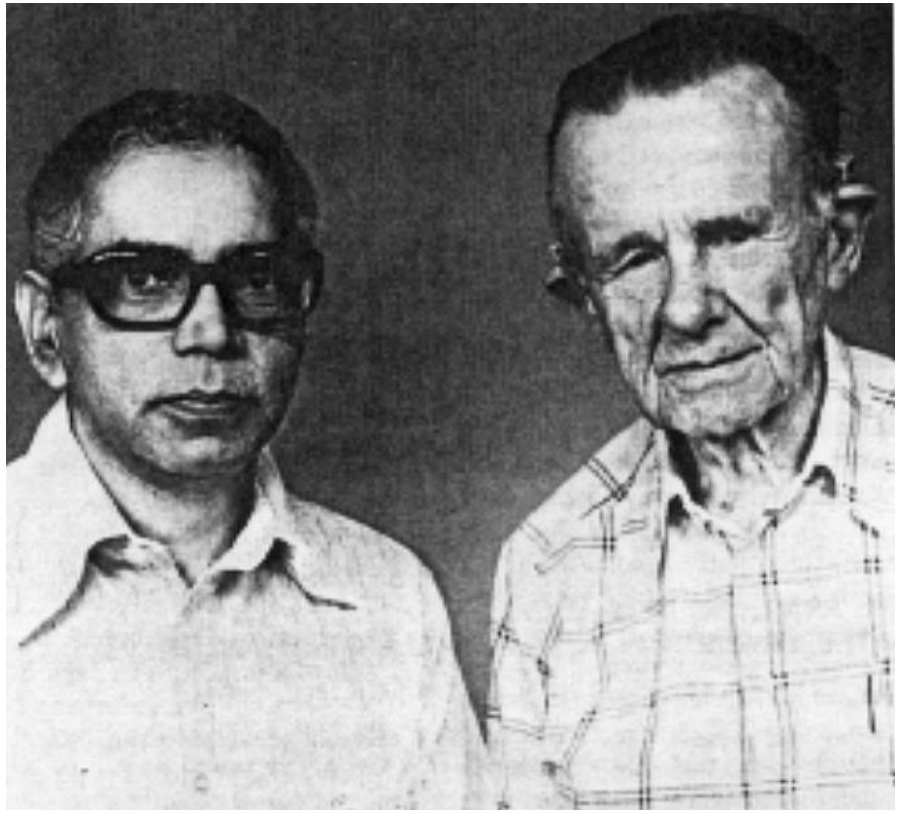
Ω_{jj} explicit expression $\sim (\Sigma^{-1})_{jj}$ optimal!

reaching semiparametric information bound

$\rightsquigarrow$ 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 $s_0 \log(p)/n$)

It is optimal!
Cramer-Rao



for data-sets with $p \approx 4'000 - 10'000$ and $n \approx 100$
$\rightsquigarrow$ 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

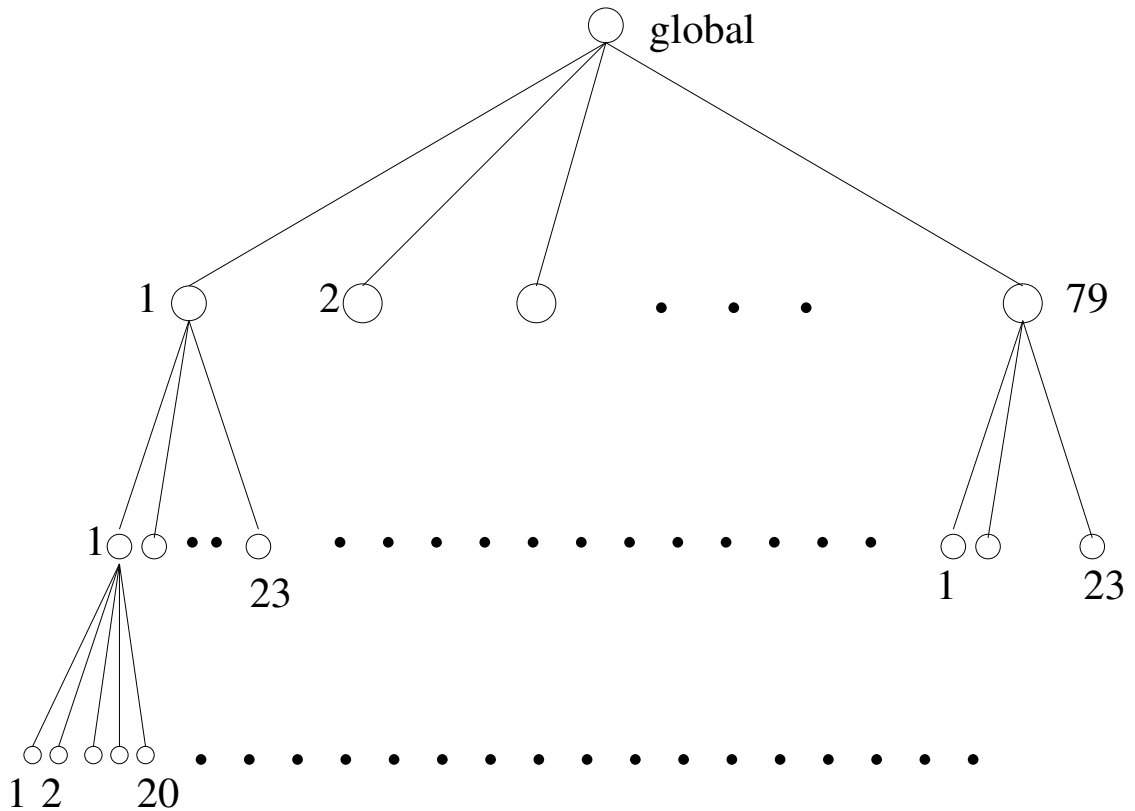
$$H_{0,jk}: \ \beta^0_{jk} = 0 \text{ versus } H_{A,jk}: \ \beta^0_{jk} \neq 0,$$
$$H_{0,G}: \ \beta^0_{jk} = 0 \text{ for all } j,k \in G \text{ versus } H_{A,G} = H^c_{0,G}$$

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

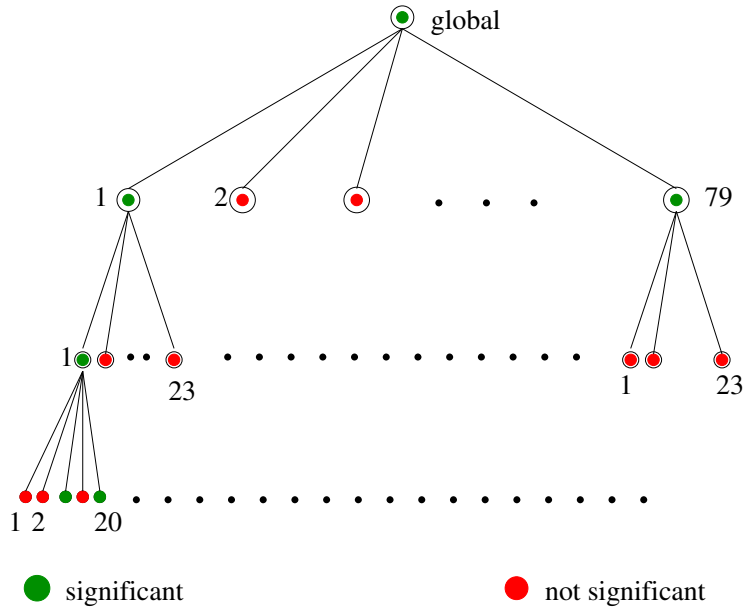
- standard: Bonferroni-Holm adjustment $\rightsquigarrow$ p-value
  $P_G \to P_{G;adj} = P_G \cdot \ell = P_G \cdot O(10^6)$!!!
- we want to do something much more efficient
  (statistically and computationally)

there is structure!

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



do hierarchical FWER adjustment (Meinshausen, 2008)

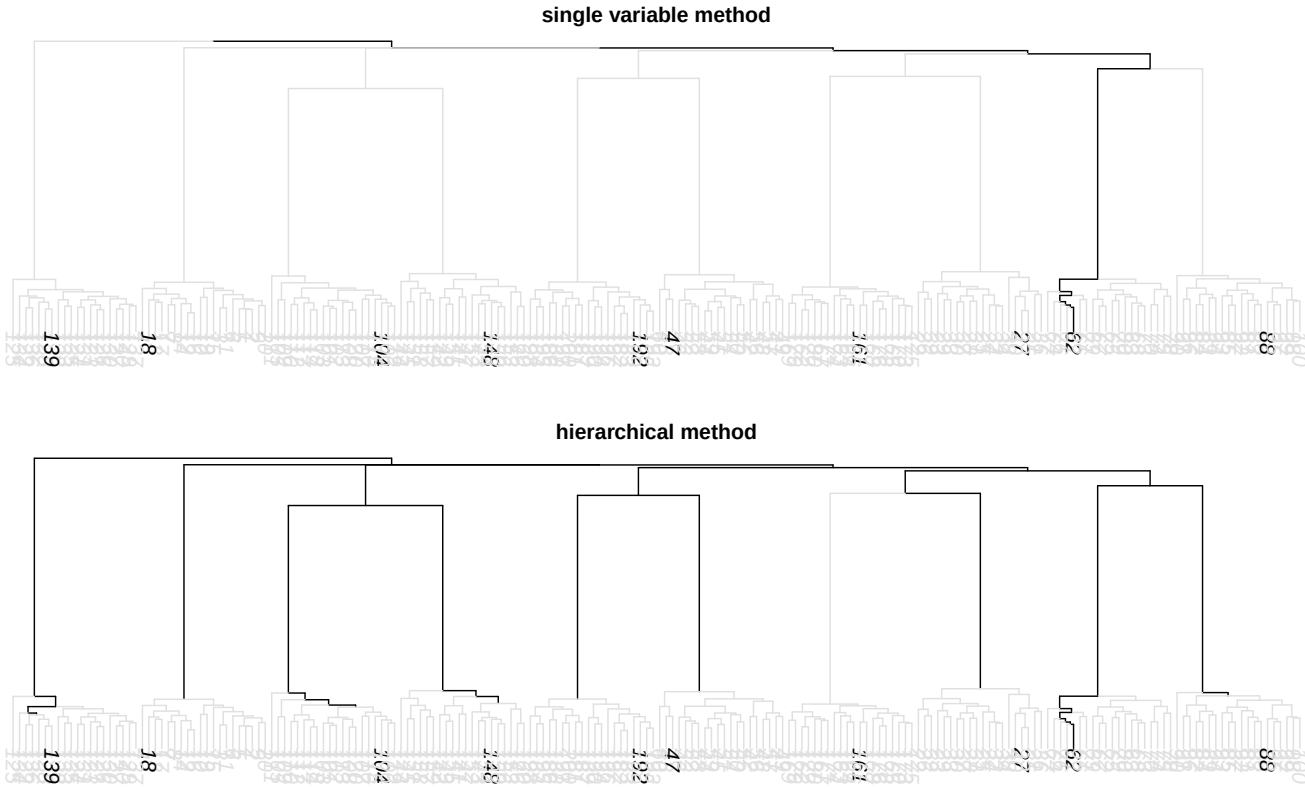


1. test global hypothesis
2. if significant: test all single response hypotheses
3. for the significant responses: test all single chromosome hyp.
4. for the significant chromosomes: test all groups of SNPs

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

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

Mandozzi & PB (2015, 2016):



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 $\rightsquigarrow$ level α
- $H_2 \leftrightarrow$ the 2 nodes of the second level of the tree
  $\rightsquigarrow$ do Bonferroni adjustment over 2 nodes
  $\rightsquigarrow$ level $\alpha/2$
- $H_3 \leftrightarrow$ the 4 nodes of the second level of the tree
  $\rightsquigarrow$ do Bonferroni adjustment over 4 nodes
  $\rightsquigarrow$ level $\alpha/4$
- ...

input:

- a hierarchy of groups/clusters $G \subseteq \{1, \ldots, p\}$
- valid p-values P_G for

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

(use de-sparsified Lasso with test-statistics $\max_{j \in G} \frac{|\hat{b}_j|}{\widehat{s.e.}_j}$)

the essential operation is very simple:

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

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

if the p-values P_G are valid, the FWER is controlled

(Meinshausen, 2008)

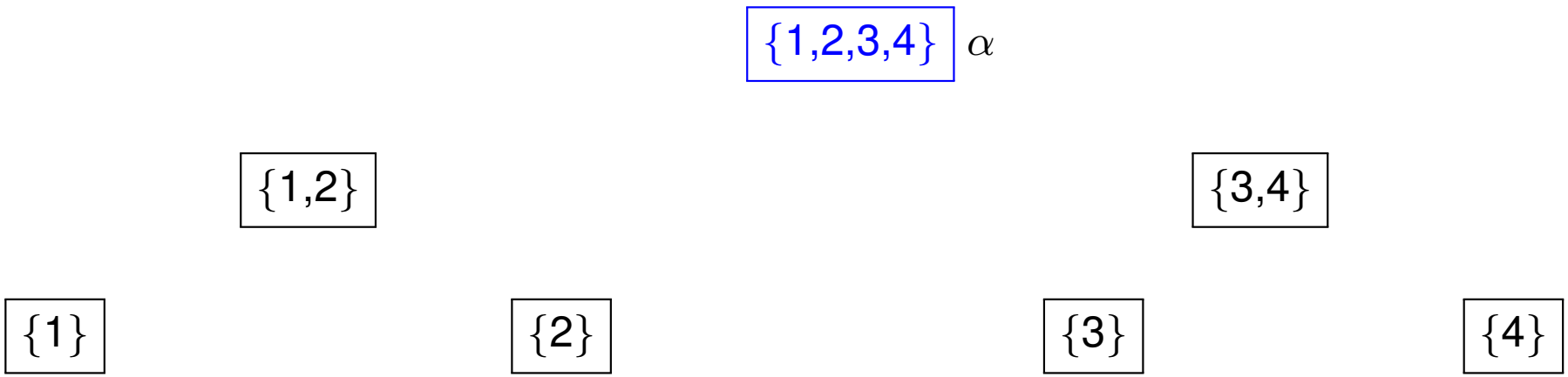
$$\Longrightarrow \mathbb{P}[\text{at least one false rejection}] \leq \alpha$$

again, for a binary tree:

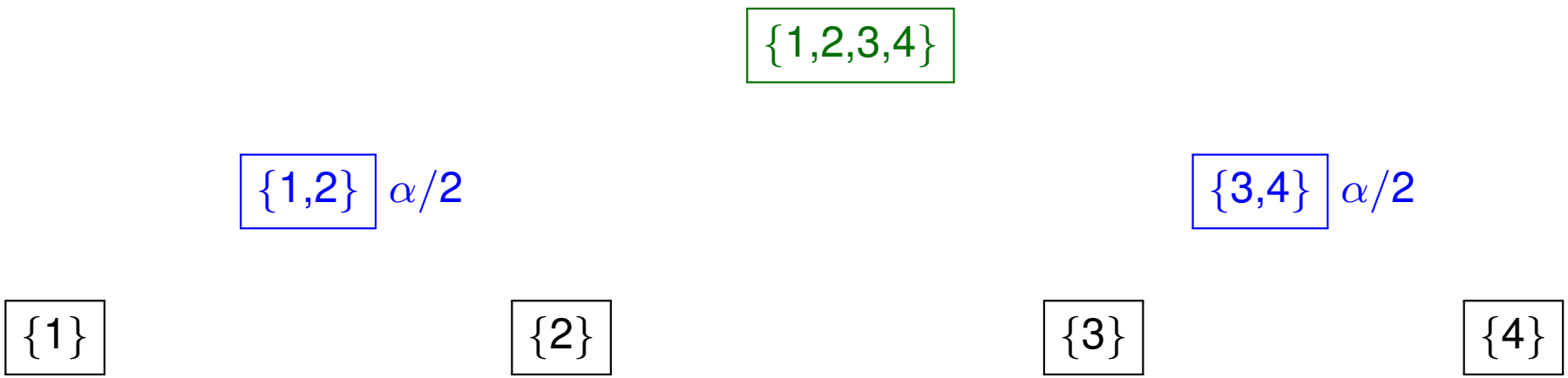
- root node: tested at level α
- next two nodes: tested at level $\approx (\alpha f_1, \alpha f_2)$ where $|G_1| = f_1 p,\ |G_2| = f_2 p$
- at a certain depth in the tree: the sum of the levels $\approx \alpha$
  on each level of depth: $\approx$ Bonferroni correction

optimizing the procedure:
α-weight distribution with inheritance (Goeman and Finos, 2012)

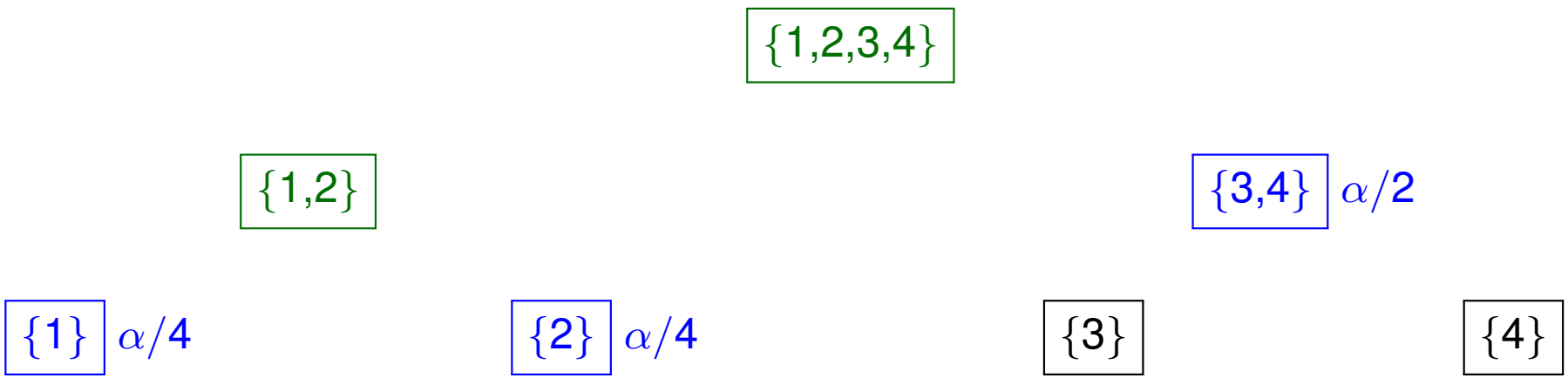
optimizing the procedure:
α-weight distribution with inheritance (Goeman and Finos, 2012)



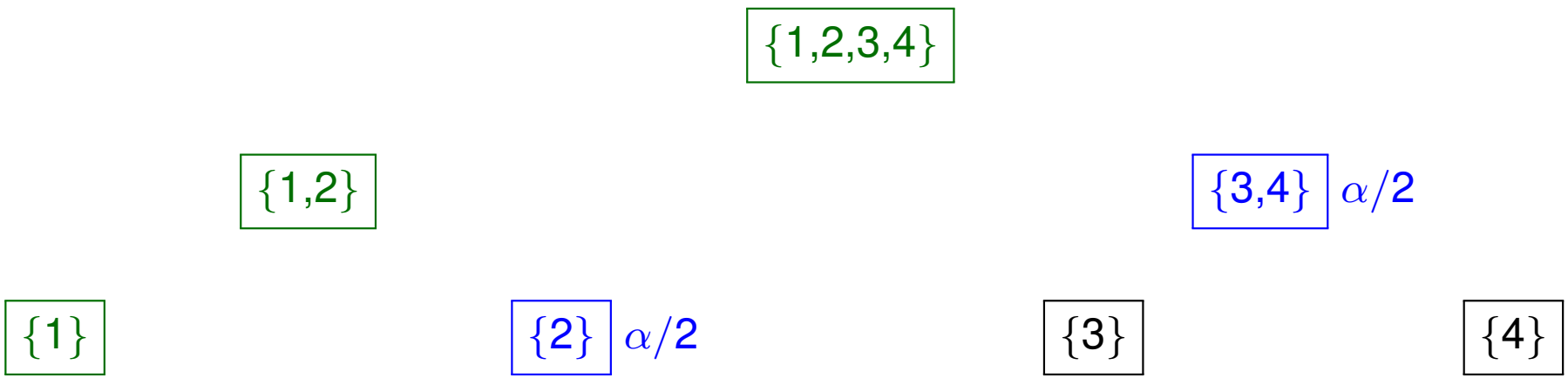
α-weight distribution with inheritance procedure (Goeman and Finos, 2012)



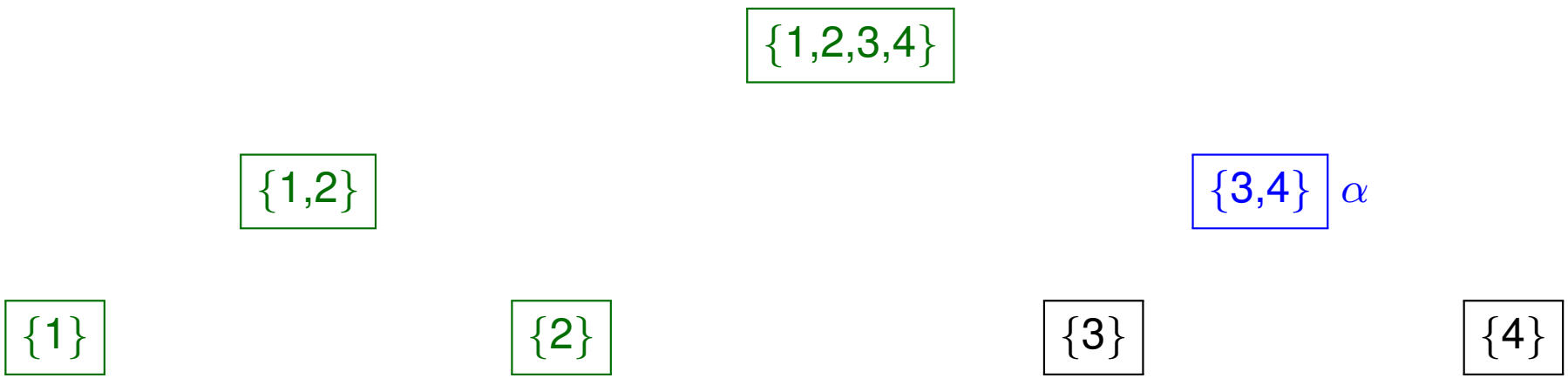
α-weight distribution with inheritance procedure (Goeman and Finos, 2012)



α-weight distribution with inheritance procedure (Goeman and Finos, 2012)

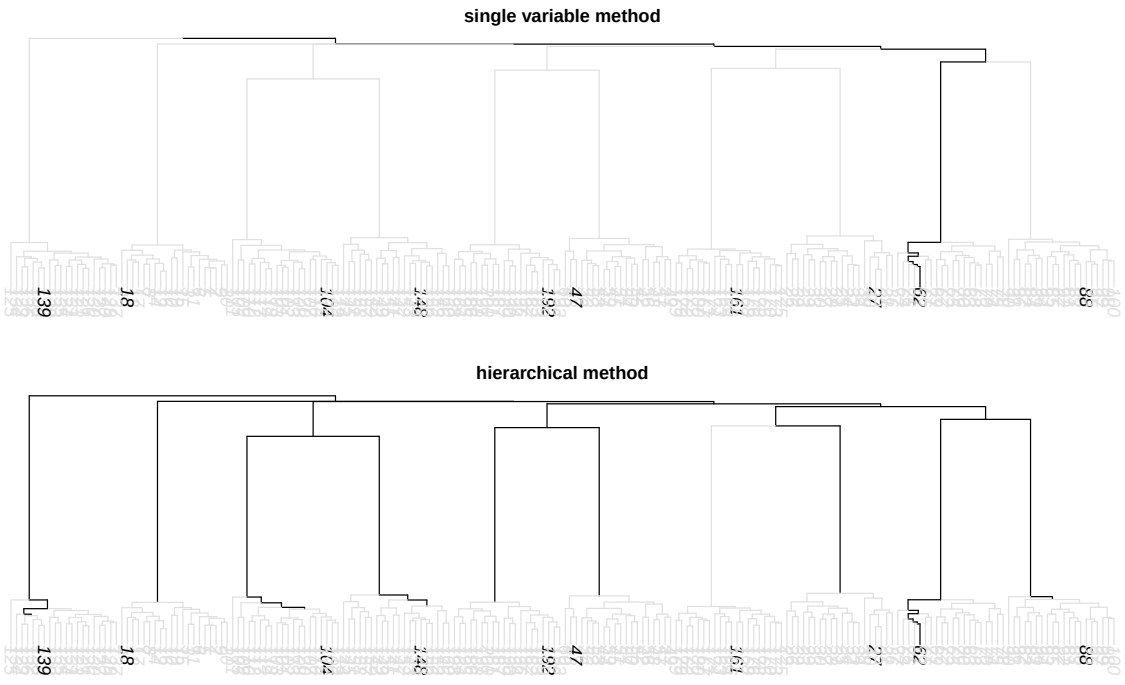


α-weight distribution with inheritance procedure (Goeman and Finos, 2012)



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

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

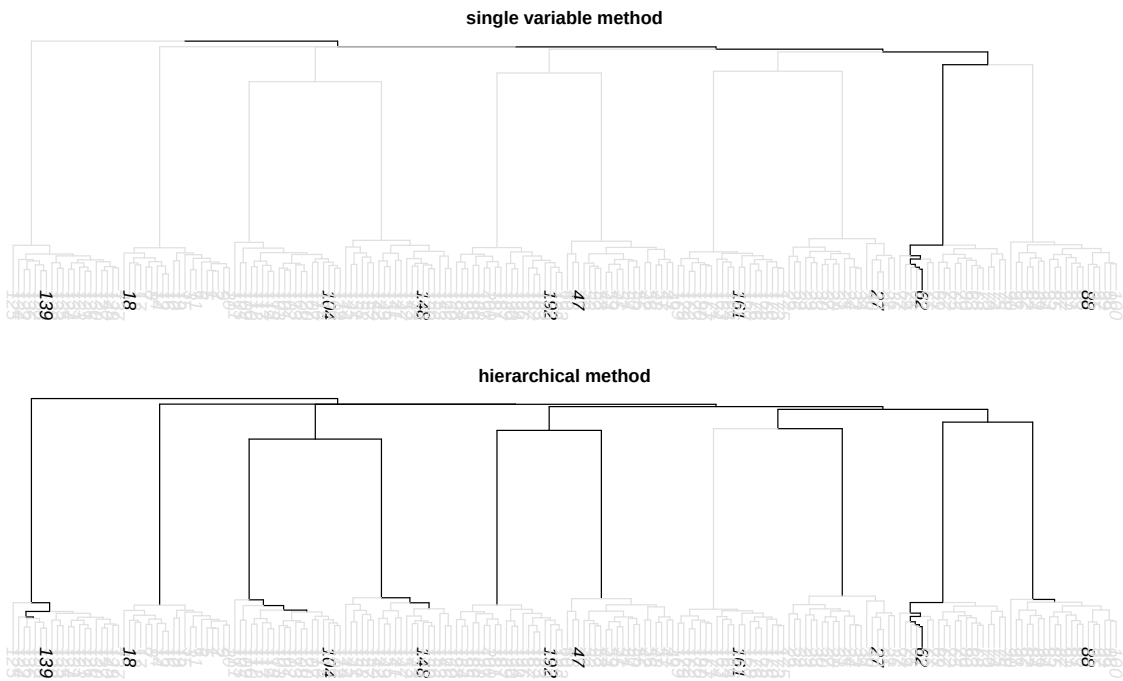


and avoid to test all possible subset of groups...!!!
which would be a disaster from a computational and multiple testing adjustment point of view

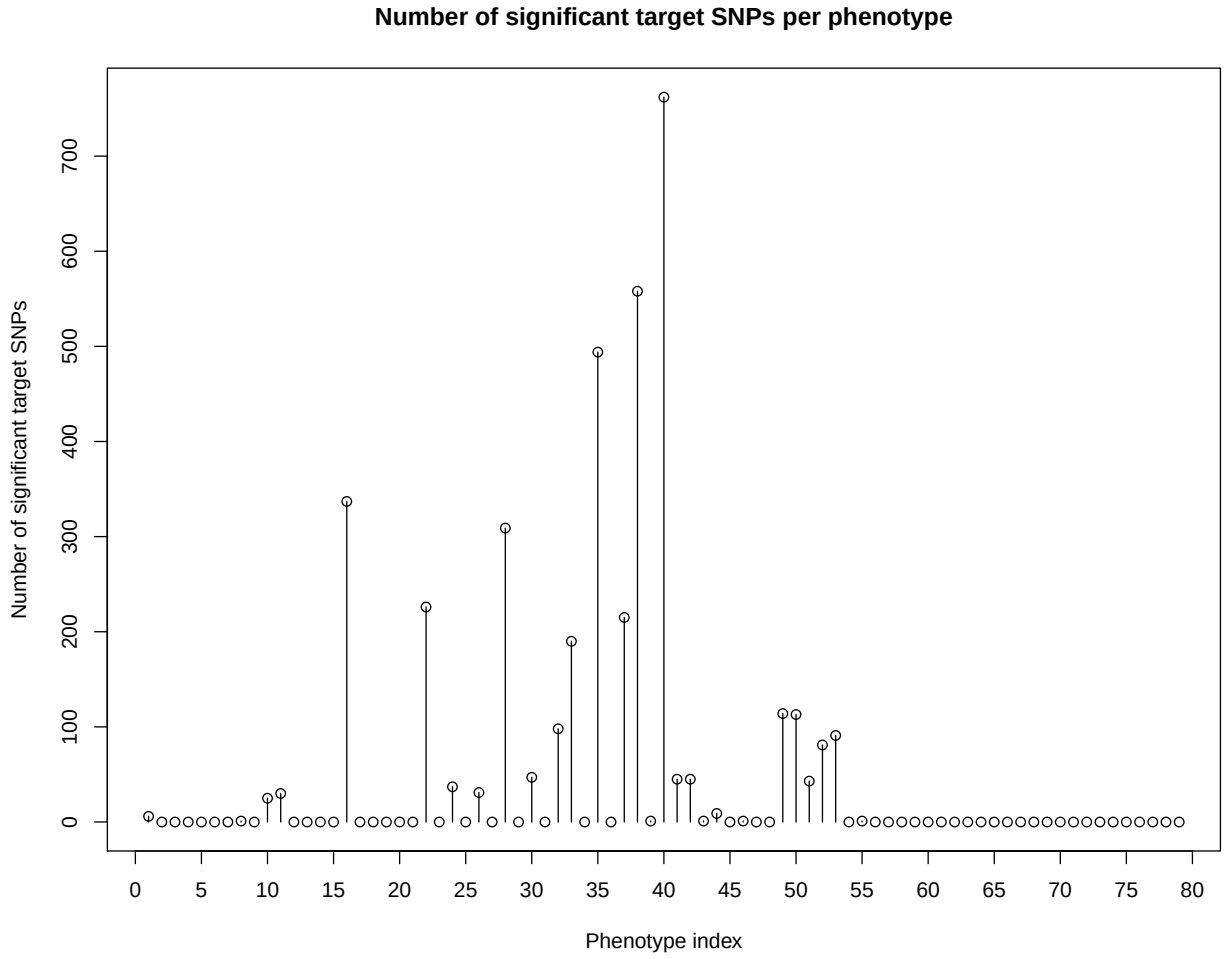
Does this work?

Mandozzi and PB (2015, 2016) provide some theory, implementation and empirical results for simulation study

- fairly reliable type I error control (control of false positives)
- reasonable power to detect true positives (and clearly better than single variable testing method)



Behavioral economics example: number of significant SNP parameters per response



response 40 (?): most significant groups of SNPs

Genomewide association studies in medicine/biology

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^{2f}
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^{2f}
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 “well-established” associations:
  - single “established” SNPs
  - small groups containing an “established” SNP

  “established”: SNP (in the group) 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
  - single or a small groups of SNPs (high resolution)
    CAD, CD, RA, T1D, T2D
  - large groups of SNPs (low resolution) only
    BD, HT

Crohn's disease

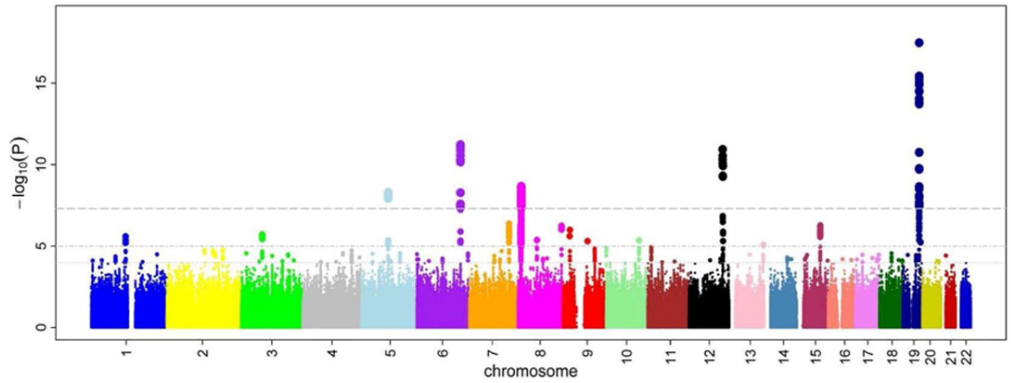
large groups

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

standard approach:
identifies single SNPs by marginal correlation



$\rightsquigarrow$ significant marginal findings cluster in regions

and then assign ad-hoc regions $\pm 10k$ base pairs around the single significant SNPs
still: this is only marginal inference

not the effect of a SNP which is adjusted by the effects of all other SNPs
i.e., not the causal SNPs

(causal direction goes from SNPs to disease status)

improvement by linear mixed models: instead of marginal correlation, try to partially adjust for presence of other SNPs (Peter Donnelly et al., Matthew Stephens et al., Peter Visscher et al.,... 2008-2016)

when adjusting for all other SNPs:

- less false positive findings!
- hierarchical inference is the “first” promising method to infer causal (groups of) SNPs

improvement by linear mixed models: instead of marginal correlation, try to partially adjust for presence of other SNPs (Peter Donnelly et al., Matthew Stephens et al., Peter Visscher et al.,... 2008-2016)

when adjusting for all other SNPs:

- less false positive findings!
- hierarchical inference is the “first” promising method to infer causal (groups of) SNPs

Genomewide association study in plant biology

push it further...

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

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

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

hierarchical inference: 4 new significant small groups
(besides nearly all known associations)

3 new associations are within and neighboring to PEPR2 gene
$\rightsquigarrow$ validation: wild-type versus pepr2-1 loss-of-function mutant
which resulted to impact root meristem
p-value = 0.0007 in Gaussian ANOVA model with 4 replicates
"a so far unknown component for root growth"

Model misspecification

true nonlinear model:

$$Y_i = f^0(X_i) + \eta_i, \ \eta_i \text{ independent of } X_i \ \ (i = 1, \ldots, n)$$

or multiplicative error

potentially heteroscedastic error:

$\mathbb{E}[\eta_i] = 0, \ \mathrm{Var}(\eta_i) = \sigma_i^2 \not\equiv \text{const.}, \eta_i' s$ independent

fitted model:

$$Y_i = X_i \beta^0 + \varepsilon_i \ \ (i = 1, \ldots, n),$$

assuming i.i.d. errors with same variances

questions:

- what is β^0 ?
- is inference machinery (uncertainty quant.) valid for β^0?

crucial conceptual difference
between random and fixed design X (when conditioning on X)

this difference is not relevant if model is true

Random design

data: n i.i.d. realizations of X
assume $\Sigma = \mathrm{Cov}(X)$ is positive definite

$$\begin{aligned} \beta^0 &= \mathrm{argmin}_\beta \mathbb{E}|f^0(X) - X\beta|^2 \qquad \text{(projection)} \\ &= \Sigma^{-1} \underbrace{(\mathrm{Cov}(f^0(X), X_1), \ldots, \mathrm{Cov}(f^0(X), X_p))^T}_{\Gamma} \end{aligned}$$

error:

$$\varepsilon = f^0(X) - X\beta^0 + \eta,$$
$$\mathbb{E}[\varepsilon|X] \neq 0, \ \mathbb{E}[\varepsilon] = 0$$

$\rightsquigarrow$ inference has to be unconditional on X

support and sparsity of β^0:

Proposition (PB and van de Geer, 2015)

$$\|\beta^0\|_r \leq (\max_{\ell} \underbrace{s_\ell}_{\ell_0\text{-spar. } X_\ell \text{ vs. } X_{-\ell}} +1)^{1/r} \|\Sigma^{-1}\|_\infty \|\Gamma\|_r \ (0 < r \leq 1)$$

If Σ exhibits block-dependence with maximal block-size $b_{\max}$:

$$\|\beta^0\|_0 \leq b_{\max}^2 |S_{f0}|$$

S_{f0} denotes the support (active) variables of $f^0(.)$

in general: linear projection is less sparse than $f^0(.)$

but ℓ_r-sparsity assump. $(0 < r \leq 1)$ is sufficient for valid inference with e.g. de-sparsified Lasso

Proposition (PB and van de Geer, 2015)

for Gaussian design: $S_0 \subseteq S_{f0}$

if a variable is significant in the misspecified linear model
$\rightsquigarrow$ it must be a relevant variable in the nonlinear function

protection against false positive findings even though the linear model is wrong

but we typically miss some true active variables

$$S_0 \overset{\text{strict}}{\subset} S_{f0}$$

message:
for random design, inference machinery for projected parameter β^0 is valid if β^0 is sparse

Proposition (PB and van de Geer, 2015)

for Gaussian design: $S_0 \subseteq S_{f^0}$

if a variable is significant in the misspecified linear model
$\rightsquigarrow$ it must be a relevant variable in the nonlinear function

protection against false positive findings even though the linear model is wrong

but we typically miss some true active variables

$$S_0 \overset{\text{strict}}{\subset} S_{f^0}$$

message:
for random design, inference machinery for projected parameter β^0 is valid if β^0 is sparse

Proposition (PB and van de Geer, 2015)

for Gaussian design: $S_0 \subseteq S_{f0}$

if a variable is significant in the misspecified linear model
$\rightsquigarrow$ it must be a relevant variable in the nonlinear function

protection against false positive findings even though the linear model is wrong

but we typically miss some true active variables

$$S_0 \overset{\text{strict}}{\subset} S_{f0}$$

message:
for random design, inference machinery for projected parameter β^0 is valid if β^0 is sparse

Fixed design (e.g. "engineering type" applications)

data: realizations of

$$Y_i = f^0(X_i) + \eta_i \ (i = 1, \ldots, n),$$

$\eta_1, \ldots, \eta_n$ independent, but potentially heteroscedastic

if $p \geq n$ and $\mathrm{rank}(X) = n$: can always write

$$f^0(X) = X\beta^0 \quad \rightsquigarrow \quad Y = X\beta^0 + \varepsilon, \ \ \varepsilon = \eta$$

for many β^0's !

take e.g. the basis pursuit solution (compressed sensing):

$$\beta^0 = \mathrm{argmin}_\beta \|\beta\|_1 \text{ such that } X\beta = (f^0(X_1), \ldots, f^0(X_n))^T$$

sparsity of β^0:
"simply" assume that there exists β^0 which is sufficiently ℓ_r-sparse $(0 < r \leq 1)$

no new theory is required

interpretation: the inference procedure leads to e.g. a confidence interval which covers all ℓ_r-sparse solutions

(PB and van de Geer, 2015)

message:
for fixed design, there is no misspecification w.r.t. linearity !
we "only" need to "bet on (weak) ℓ_r-sparsity"

Further issues

the bootstrap: more reliable and powerful inference
$\rightsquigarrow$ better finite-sample approximation (empirically) and more powerful multiple testing correction under dependence

the work from the 1980's can be used in the modern context of high-dimensional inference!

computation:
the de-sparsified Lasso has $O(p^2 n^2)$ computational cost

work in progress to improve this

Conclusions

key concepts for high-dimensional statistics:

- sparsity of the underlying regression vector
  - sparse estimator is optimal for prediction
  - non-sparse estimators are optimal for uncertainty quantification
- identifiability via restricted eigenvalue assumption

hierarchical inference:

- very powerful to detect significant groups of variables at data-driven resolution
- exhibits impressive performance and validation on bio-/medical data

model misspecification: some issues have been addressed (PB & van de Geer, 2015)

bootstrapping non-sparse estimators improves inference (Dezeure, PB & Zhang, 2016)

robustness, reliability and reproducibility of results...

in view of (yet) uncheckable assumptions

$\rightsquigarrow$

confirmatory high-dimensional inference
remains an interesting challenge

Thank you!

robustness, reliability and reproducibility of results...

in view of (yet) uncheckable assumptions

$\rightsquigarrow$

confirmatory high-dimensional inference
remains an interesting challenge

Thank you!

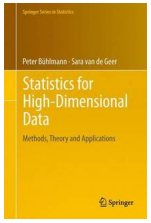
Software:

`R-package hdi` (Meier, Dezeure, Meinshausen, Mächler & PB, since 2013)

`Bioconductor-package hierGWAS` (Buzdugan, 2016)

References to some of our own work:

- Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methodology, Theory and Applications. Springer.



- van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics 42, 1166-1202.
- Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. Electronic Journal of Statistics 9, 1449-1473.
- Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R-software hdi. Statistical Science 30, 533–558.
- Mandozzi, J. and Bühlmann, P. (2016). Hierarchical testing in the high-dimensional setting with correlated variables. Journal of the American Statistical Association 111, 331-343.
- Mandozzi, J. and Bühlmann, P. (2015). A sequential rejection testing method for high-dimensional regression with correlated variables. International Journal of Biostatistics 12, 79-95.
- Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E. and Bühlmann, P. (2016). Assessing statistical significance in joint analysis for genome-wide association studies. Bioinformatics 32, 1990-2000.
- Klasen, J.R., Barbez, E., Meier, L., Meinshausen, N., Bühlmann, P., Koornneef, M., Busch, W. and Schneeberger, K. (2016). A multi-marker association method for Genome-Wide Association studies without the need for population structure correction. Nature Communications 7, Article number 13299 (2016).
- Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2016). High-dimensional simultaneous inference with the bootstrap. To appear in TEST, with discussion.

Computational issue

de-sparsified Lasso for all components $j = 1, \ldots, p$:

requires $p + 1$ Lasso regressions
for $p \gg n$: $O(p^2 n^2)$ computational cost

$p = O(10^6) \rightsquigarrow O(10^{12} n^2)$ despite trivial distributed computing

work in progress with Rajen Shah using thresholded Ridge or generalized LS

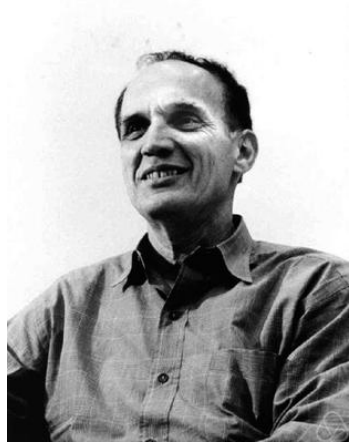
the GWAS examples have been computed with preliminary Lasso variable screening and multiple sample splitting

The bootstrap (Efron, 1979): more reliable inference

residual bootstrap for fixed design:

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

$\hat{\varepsilon} = Y - X\hat{\beta}$, $\hat{\beta}$ from the Lasso


Efron

- i.i.d. resampling of centered residuals $\hat{\varepsilon}_i \rightsquigarrow \varepsilon_1^*, \ldots, \varepsilon_n^*$
- wild bootstrapping for heteroscedastic errors (Wu (1986), Mammen (1993)):

$\varepsilon_i^* = W_i \hat{\varepsilon}_i$, $W_1, \ldots, W_n$ i.i.d. $\mathbb{E}[W_i] = \mathbb{E}[W_i^3] = 0$

then:

$$Y^* = X\hat{\beta} + \varepsilon^*$$

bootstrap sample: $(X_1, Y_1^*), \ldots, (X_n, Y_n^*)$

goal: distribution of an algorithm/estimator $\hat{\theta} = g(\{X_i, Y_i\}_{i=1}^n)$

goal: distribution of an algorithm/estimator $\hat{\theta} = g(\{X_i, Y_i\}_{i=1}^n)$

compute algorithm/estimator

$$\hat{\theta}^* = g(\underbrace{\{X_i, Y_i^*\}_{i=1}^n}_{\text{bootstrap sample}}) \text{ (plug-in principle)}$$

many times to approximate the true distribution of $\hat{\theta}$

(with importance sampling for some cases...)

bootstrapping the Lasso $\rightsquigarrow$ "bad" because of sparsity of the estimator and super-efficiency phenomenon


Joe Hodges

- poor for estimating uncertainty about non-zero regression parameters
- uncertainty about zero parameters overly optimistic

one should bootstrap a regular non-sparse estimator

(Giné & Zinn, 1989, 1990)

$\rightsquigarrow$ bootstrap the de-sparsified Lasso $\hat{b}$

(Dezeure, PB & Zhang, 2016)

Bootstrapping the de-sparsified Lasso (Dezeure, PB & Zhang, 2016)

assumptions:

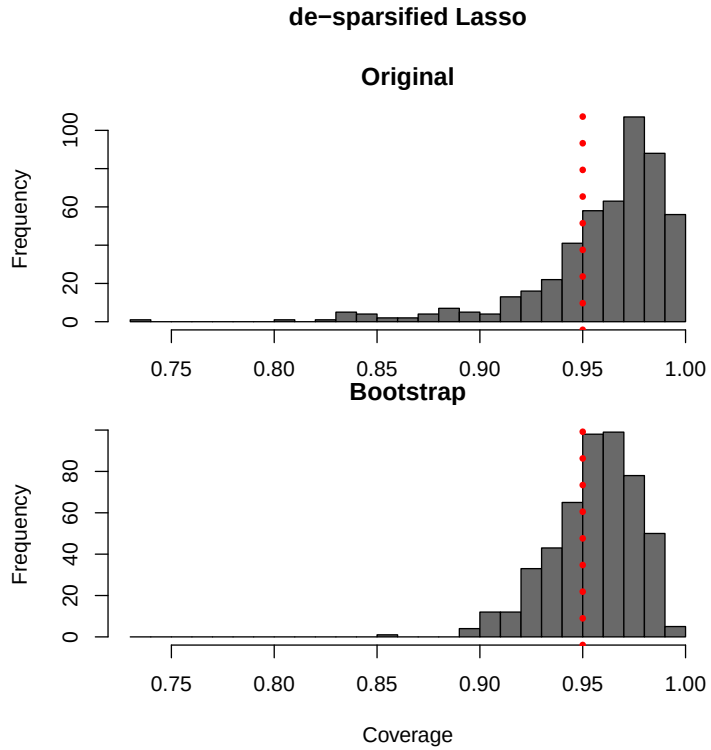
- linear model with fixed design $Y = X\beta^0 + \varepsilon$ "always true"
- sparsity for Y vs. X: $s_0 = o(n^{1/2}\log(p)^{-3/2})$ "OK"
  sparsity X_j vs. X_{-j} real assumption
- errors can be heteroscedastic and non-Gaussian with 4th moments (wild bootstrap for heter. errors) weak assumption
- $\log(p)^7 = o(n)$ weak assumption

$\rightsquigarrow$ consistency of the bootstrap for simultaneous inference!

$$\sup_c \left| \mathbb{P}[\max_{j=1,\ldots,p} \pm \frac{\hat{b}_j - \beta_j^0}{\widehat{s.e.}_j} \leq c] - \mathbb{P}^*[\max_{j=1,\ldots,p} \pm \frac{\hat{b}_j^* - \hat{\beta}_j}{\widehat{s.e.}_j^*} \leq c] \right| = o_P(1)$$

(Dezeure, PB & Zhang, 2016)

involves very high-dimensional maxima of non-Gaussian (but limiting Gaussian) quantities (cf. Chernozhukov et al. (2013))



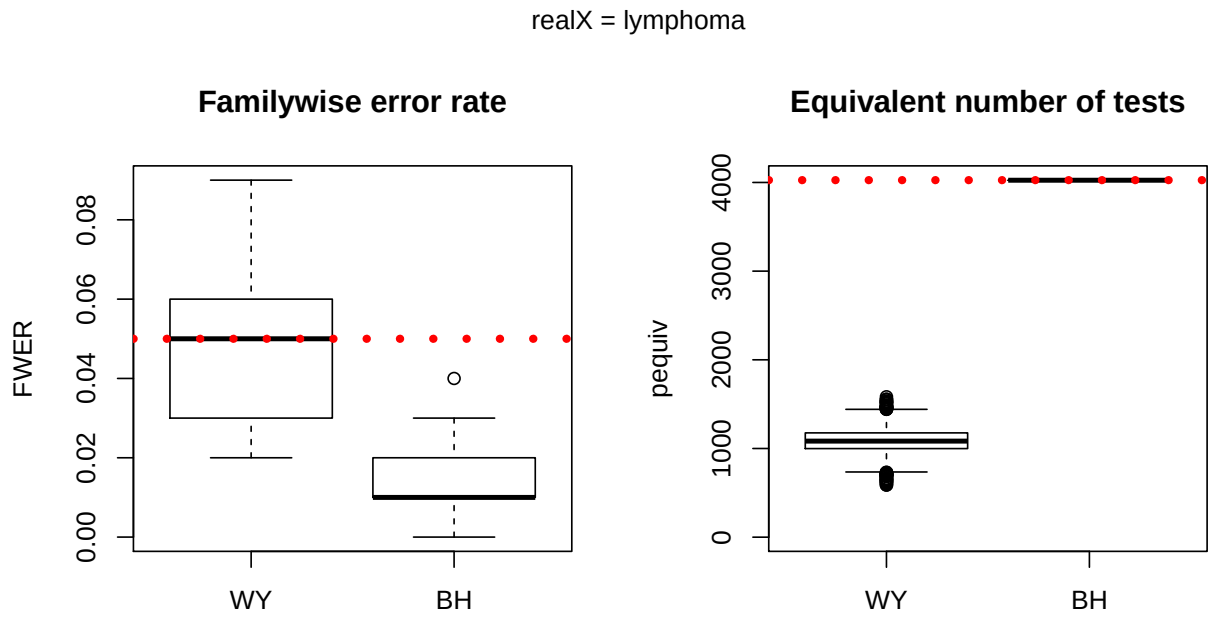
implications:

- more reliable confidence intervals and tests for individual parameters
- powerful simultaneous inference for many parameters
- more powerful multiple testing correction (than Bonferroni-Holm), in spirit of Westfall and Young (1993): effective dimension is e.g. $p_{\text{eff}} = 100K$ instead of $p = 1M$

this seems to be the “state of the art” technique at the moment

more powerful multiple testing correction (than Bonferroni-Holm)

effective dimension is e.g. $p_{\text{eff}} \approx 1000$ instead of $p \approx 4000$

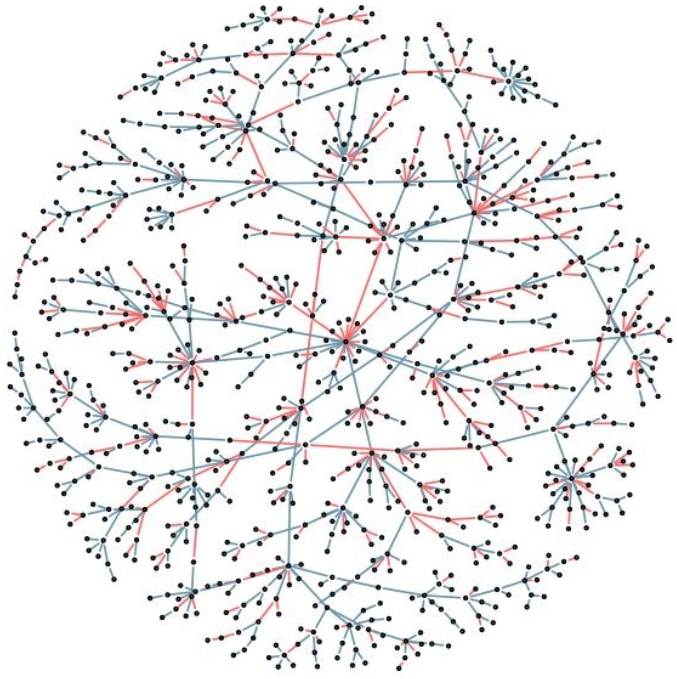


need to control under the "complete null-hypotheses"

$$\mathbb{P}[\max_{j=1,\ldots,p} |\hat{b}_j / \widehat{s.e.}_j| \leq c] \approx \mathbb{P}^*[\max_{j=1,\ldots,p} |\hat{b}_j^* / \widehat{s.e.}_j^*| \leq c]$$

maximum over (highly) correlated components with p variables is equivalent to maximum of p_{eff} independent components

Outlook: Network models



Gaussian Graphical model
Ising model

undirected edge encodes conditional dependence given all other random variables

problem: given data, infer the undirected edges
Gaussian Graphical model: (Meinshausen & PB, 2006)
Ising model: (Ravikumar, Wainwright & Lafferty; 2010)

$\rightsquigarrow$ uncertainty quantification; "similarly" as discussed