P-values and confidence intervals for high-dimensional problems

Peter Bühlmann

Seminar für Statistik, ETH Zürich

June 2015

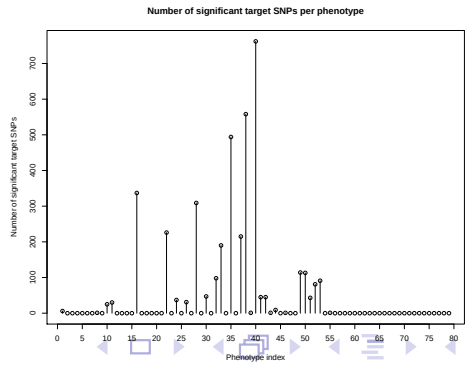
High-dimensional data

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 editor's 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 editor's discretion and at the referees' suggestion."
- "Too few budding scientists receive adequate training in statistics and other quantitative aspects of their subject."

statistics is important...

and its mathematical roots as well !

statistics is important...

and its mathematical roots as well !

P-values for high-dimensional linear models

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

want uncertainty quantification!

goal: statistical hypothesis testing

$$H_{0,j} : \beta_j^0 = 0 \quad \text{or} \quad H_{0,G} : \beta_j^0 = 0 \text{ for all } j \in G \subseteq \{1, \ldots, p\}$$

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 $n \to \infty$

P-values for high-dimensional linear models

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

want uncertainty quantification!

goal: statistical hypothesis testing

$$H_{0,j} : \beta_j^0 = 0 \quad \text{or} \quad H_{0,G} : \ \beta_j^0 = 0 \text{ for all } j \in G \subseteq \{1, \ldots, p\}$$

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 $n \to \infty$

$\rightsquigarrow$ standard bootstrapping and subsampling cannot be used

Low-dimensional projections and bias correction

Or de-sparsifying the Lasso estimator

related work by Zhang and Zhang (2011; publ. 2014)

motivation:

$\hat{\beta}_{\mathrm{LS},j}$ 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:

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

$\hat{b}_j$ is not sparse!... and this is crucial to obtain Gaussian limit nevertheless: it is "optimal" (see later)

- 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)

$\hat{b}_j$ is not sparse!... and this is crucial to obtain Gaussian limit nevertheless: it is "optimal" (see later)

- 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, 2013)

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 \Omega_{jj}) \quad (j = 1, \ldots, p)$$

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 !

Asymptotic pivot and optimality

Theorem (van de Geer, PB, Ritov & Dezeure, 2013)

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 \Omega_{jj}) \quad (j = 1, \ldots, p)$$

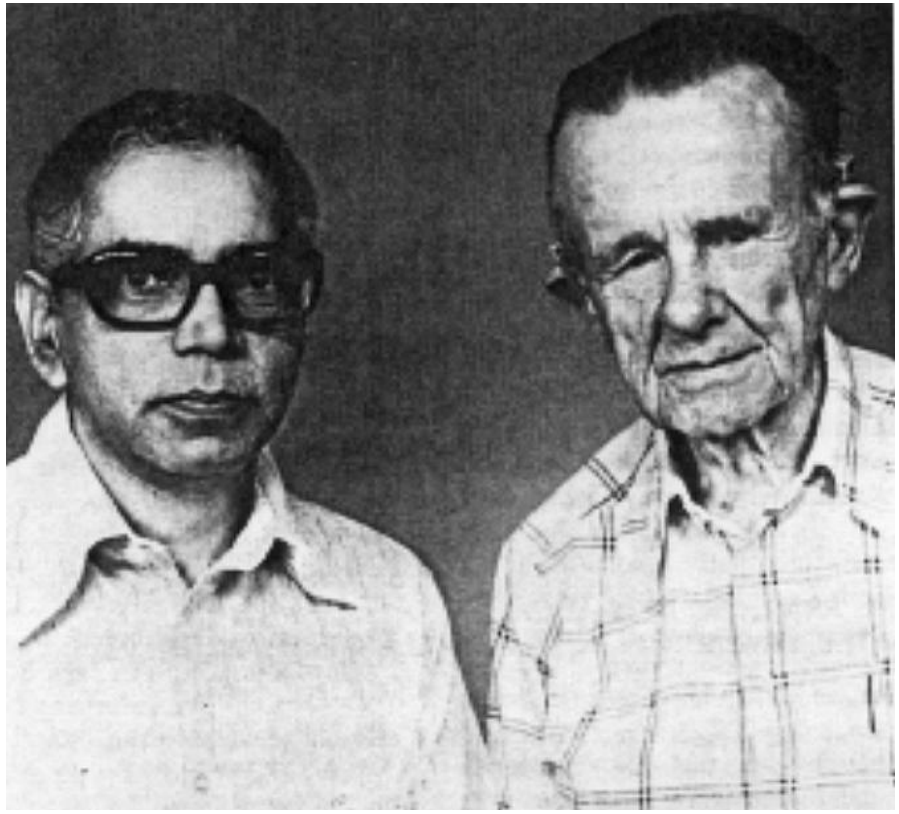
Ω_{jj} explicit expression $\sim (\Sigma^{-1})_{jj}$ optimal!

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 !

It is optimal!
Cramer-Rao



Uniform convergence:

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 \Omega_{jj}) \quad (j = 1, \ldots, p)$$

convergence is uniform over $\mathcal{B}(s_0) = \{\beta; \|\beta\|_0^0 \leq s_0\}$

$\rightsquigarrow$ honest tests and confidence regions!

and we can avoid post model selection inference
(cf. Pötscher and Leeb)

Simultaneous inference over all components:

$$\sqrt{n}(\hat{b} - \beta^0) \approx (W_1, \ldots, W_p) \sim \mathcal{N}_p(0, \sigma_\varepsilon^2 \Omega)$$

$\rightsquigarrow$ can construct P-values for:

$H_{0,G}$ with any G: test-statistics $\max_{j \in G} |\hat{b}_j|$
since covariance structure Ω is known

and
can easily do efficient multiple testing adjustment since
covariance structure Ω is known!

Alternatives?

- versions of bootstrapping (Chatterjee & Lahiri, 2013)

  $\rightsquigarrow$ super-efficiency phenomenon!
  i.e. non-uniform convergence


Joe Hodges

  - good for estimating the zeroes (i.e., $j \in S_0^c$ with $\beta_j^0 = 0$)
  - bad for estim. the non-zeroes (i.e., $j \in S_0$ with $\beta_j^0 \neq 0$)

- multiple sample splitting (Meinshausen, Meier & PB, 2009)
  split the sample repeatedly in two halfs:
  - select variables on first half
  - p-values using second half, based on selected variables

  $\rightsquigarrow$ avoids (because of sample splitting) over-optmistic p-values, but potentially suffers in terms of power

Some further remarks on multiple sample splitting

- if the (generalized linear) model is correct:
  it "works" for fixed and random design
- in misspecified models:
  it "works" for random design for the "best projected parameter" (see later)

the theoretical justification assumes the variable screening property:

$$\underbrace{\hat{S}}_{\text{based on 1st half-sample}} \supseteq S_0$$

(or a slightly relaxed form (PB and Mandozzi, 2014))
$\rightsquigarrow$ not nice...

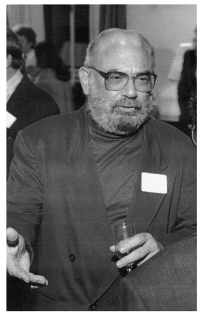
but: the method performs rather well in broad simulation study (Dezeure, PB, Meier and Meinshausen, 2014)

... the method performs rather well in broad simulation study
the heuristic reason:

- B sample splits: p-values $P_j^{(1)}, \ldots, P_j^{(B)}$ for $H_{0,j} : \beta_j^0 = 0$

$$P_j^{(b)} = \begin{cases} 1 & \text{if } j \notin \hat{S}^{(b)} \\ \text{p-val from t-test on 2nd half-sample} & \text{if } j \in \hat{S}^{(b)}. \end{cases}$$

- need to aggregate these dependent p-values



Leo Breiman

a simple rule (Meinshausen, Meier and PB, 2009)

$$P_j^{(\text{aggr})} = \text{sample-median}(2P_j^{(1)}, \ldots, 2P_j^{(B)})$$

$P_j^{\text{aggr}} < 1 \iff$ variable j has been selected in $> 50\%$ of the B sample splits

$\rightsquigarrow$ an important stability property
the method is conservative

First real data results

where we have collaborated in joint projects

- Motif regression (computational biology)
  $n = 143,\ p = 196$

  with desparsified Lasso and multiple sample splitting:
  one significant single variable at 5% level with FWER multiple testing adjustment
- Riboflavin production with Bacillus Subtilis (genomics)
  $n = 71,\ p = 4096$

  with desparsified Lasso and multiple sample splitting:
  one significant single variable at 5% level with FWER multiple testing adjustment

surprising?

remember the meaning of β_j^0:

it measures effect which is adjusted for by all other variables...

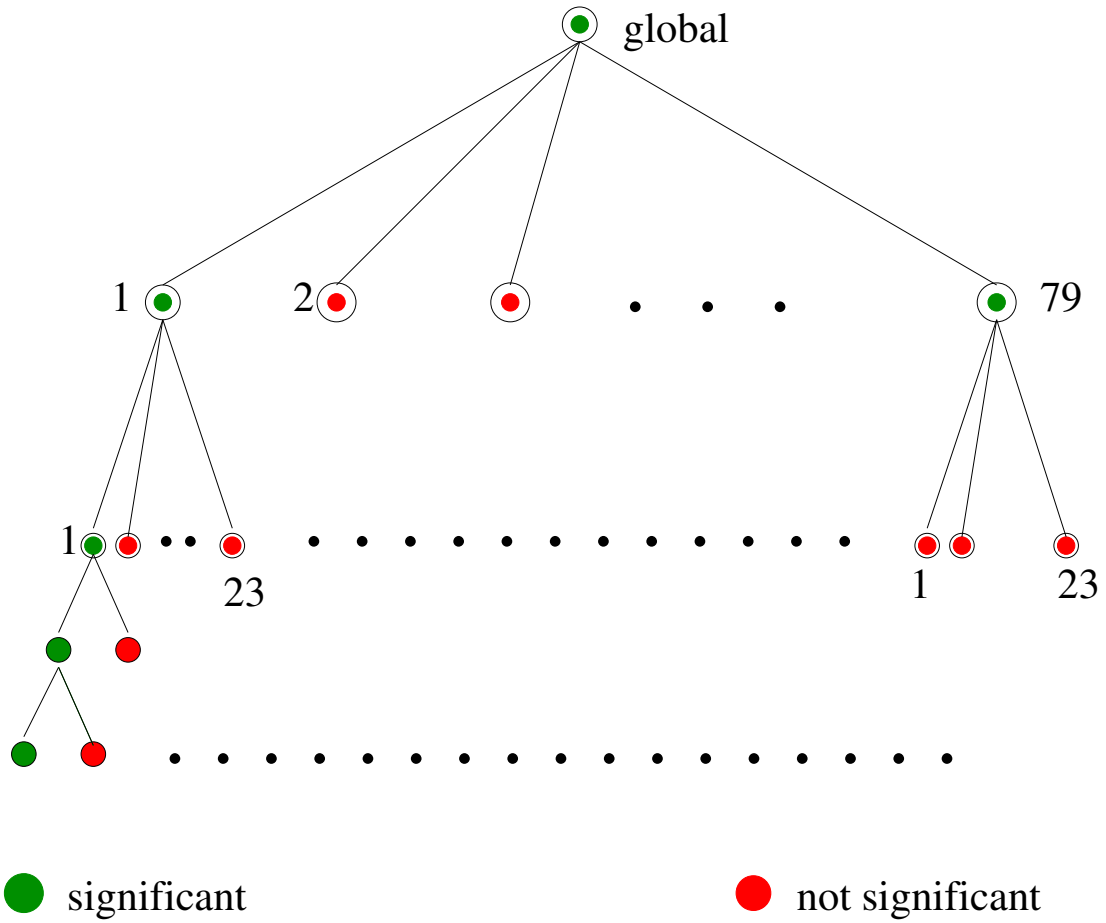
Behavioral economics and genetics (with Ernst Fehr, U. Zurich)
$n = 1'525,\ p \approx 0.5 \cdot 10^6$
(and 79 response variables, measuring “behavior”)

$\rightsquigarrow$ cannot detect any single variable as significant after standard multiple testing correction

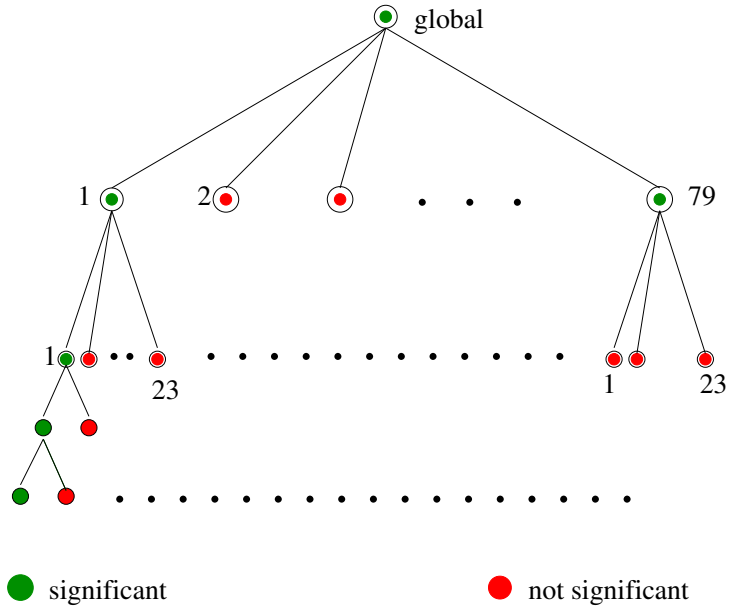
Hierarchical inference

there is structure!

- 79 response experiments
- 23 chromosomes per response experiment
- groups from hierarchical clustering 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 finer groups from hierarchical clustering

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

input:

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

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

output:
p-values for groups/clusters which control the familyw. err. rate (FWER = $\mathbb{P}$[at least one false positive/rejection])
with hierarchical constraints:

if $H_{0,G}$ is not rejected

$\Longrightarrow$ $H_{0,\tilde{G}}$ not rejected for $\tilde{G}$ lower in the hierarchy/tree

see Meinshausen (2008)
and for general sequential testing principle (Goeman and Solari, 2010)

the essential operation is very simple:

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

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

- 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

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

(Meinshausen, 2008)

$$\text{reject } H_{0,G} \text{ if } P_{G;\mathrm{hier-adj}} \leq \alpha$$
$$\Longrightarrow \quad \mathbb{P}[\text{at least one false rejection}] \leq \alpha$$

optimized procedure:

- using Shaffer's improvement
  exploiting logical relations among hypotheses:
  if $H_{0,G}$ is true, all $H_{0,G'}$ are true for $G' \subseteq G$
- using additional sequential-type testing principles
  (aka Bonferroni-Holm instead of Bonferroni)

Bonferroni-Holm

Hypotheses to be tested: $\{1\}$ $\{2\}$

1st step:

adjusted p-values : $2P_{\{1\}}$ $2P_{\{2\}}$

FWER control (no false rejection at all): $\alpha/2 + \alpha/2 = \alpha$

If one null hypothesis (e.g. $H_{\{1\}}$) is rejected:
do 2nd step with improved multiplicity: $P_{\{2\}}$

optimized procedure:

- using Shaffer's improvement
  exploiting logical relations among hypotheses:
  if $H_{0,G}$ is true, all $H_{0,G'}$ are true for $G' \subseteq G$
- using additional sequential-type testing principles
  (aka Bonferroni-Holm instead of Bonferroni)

Bonferroni-Holm

Hypotheses to be tested: $\boxed{\{1\}}$ $\boxed{\{2\}}$

1st step:

adjusted p-values : $2P_{\{1\}}$ $2P_{\{2\}}$

FWER control (no false rejection at all): $\alpha/2 + \alpha/2 = \alpha$

If one null hypothesis (e.g. $H_{\{1\}}$) is rejected:
do 2nd step with improved multiplicity: $P_{\{2\}}$

optimized procedure:

- using Shaffer's improvement
  exploiting logical relations among hypotheses:
  if $H_{0,G}$ is true, all $H_{0,G'}$ are true for $G' \subseteq G$
- using additional sequential-type testing principles
  (aka Bonferroni-Holm instead of Bonferroni)

Bonferroni-Holm

Hypotheses to be tested: $\boxed{\{1\}}$ $\boxed{\{2\}}$

1st step:

adjusted p-values : $2P_{\{1\}}$ $2P_{\{2\}}$

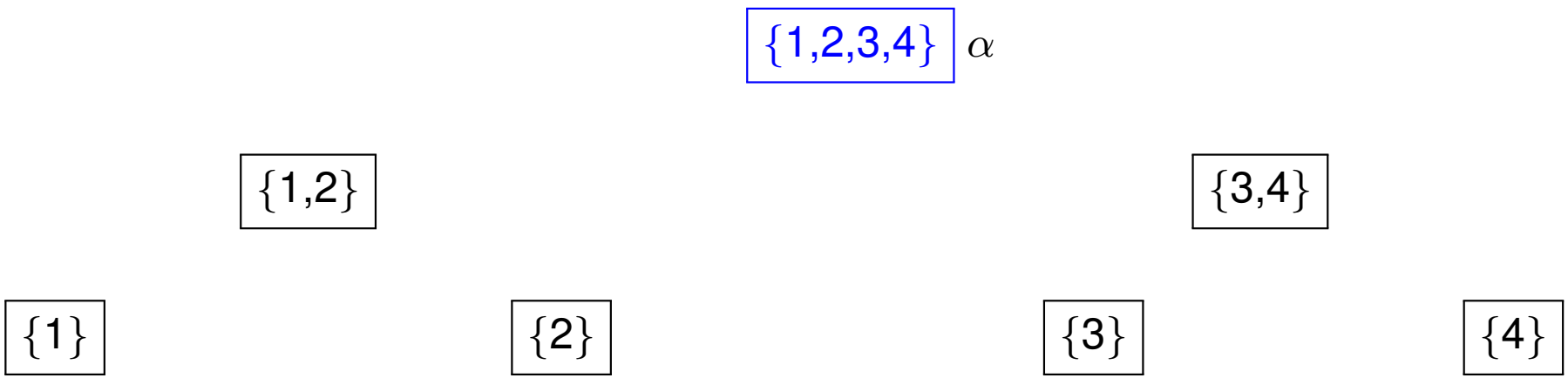
FWER control (no false rejection at all): $\alpha/2 + \alpha/2 = \alpha$

If one null hypothesis (e.g. $H_{\{1\}}$) is rejected:
do 2nd step with improved multiplicity: $P_{\{2\}}$

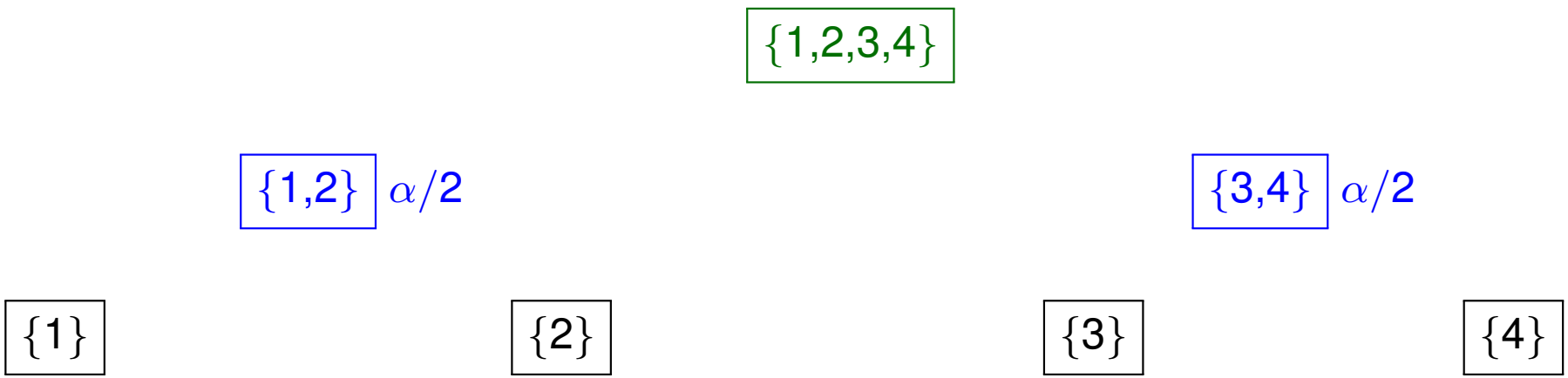
α-weight distribution with inheritance procedure (Goeman and Finos, 2012)

α-weight distribution with inheritance procedure

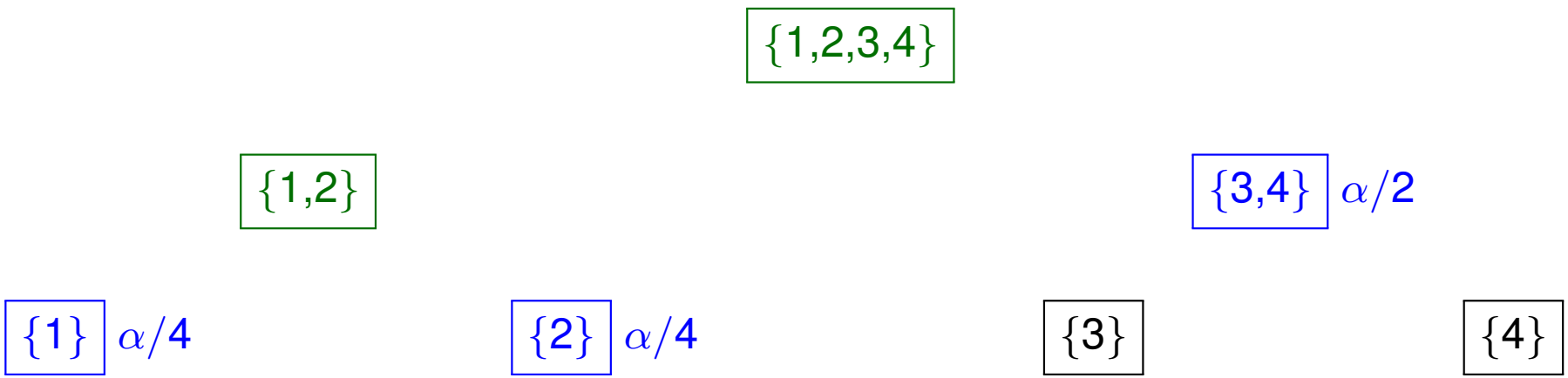
(Goeman and Finos, 2012)



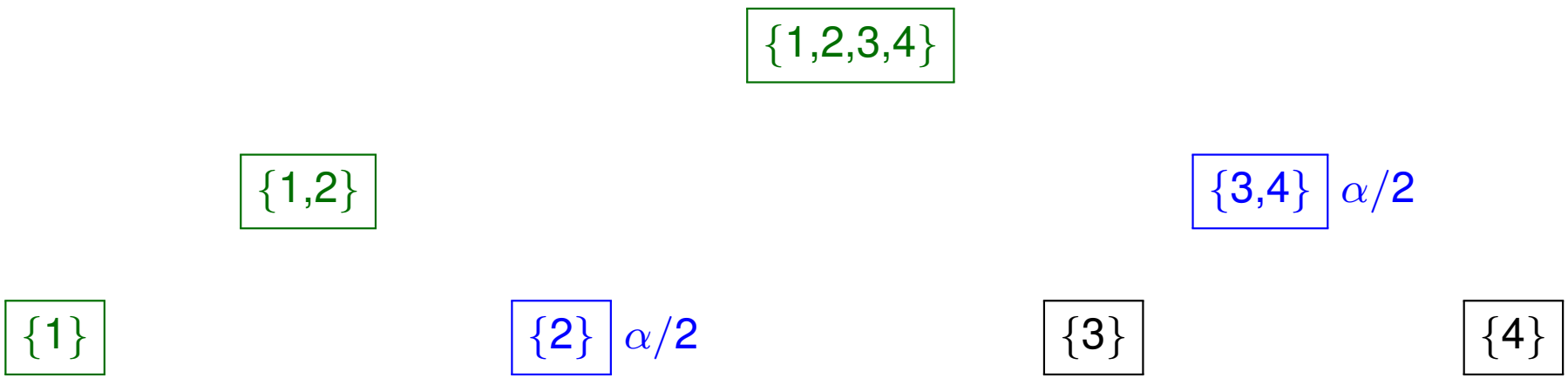
α-weight distribution with inheritance procedure (Goeman and Finos, 2012)



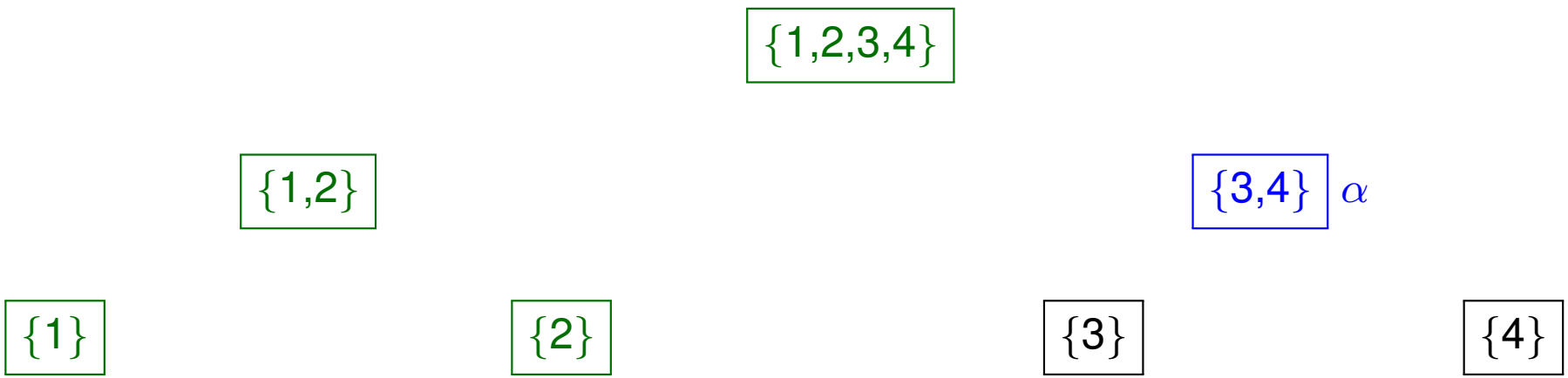
α-weight distribution with inheritance procedure (Goeman and Finos, 2012)



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

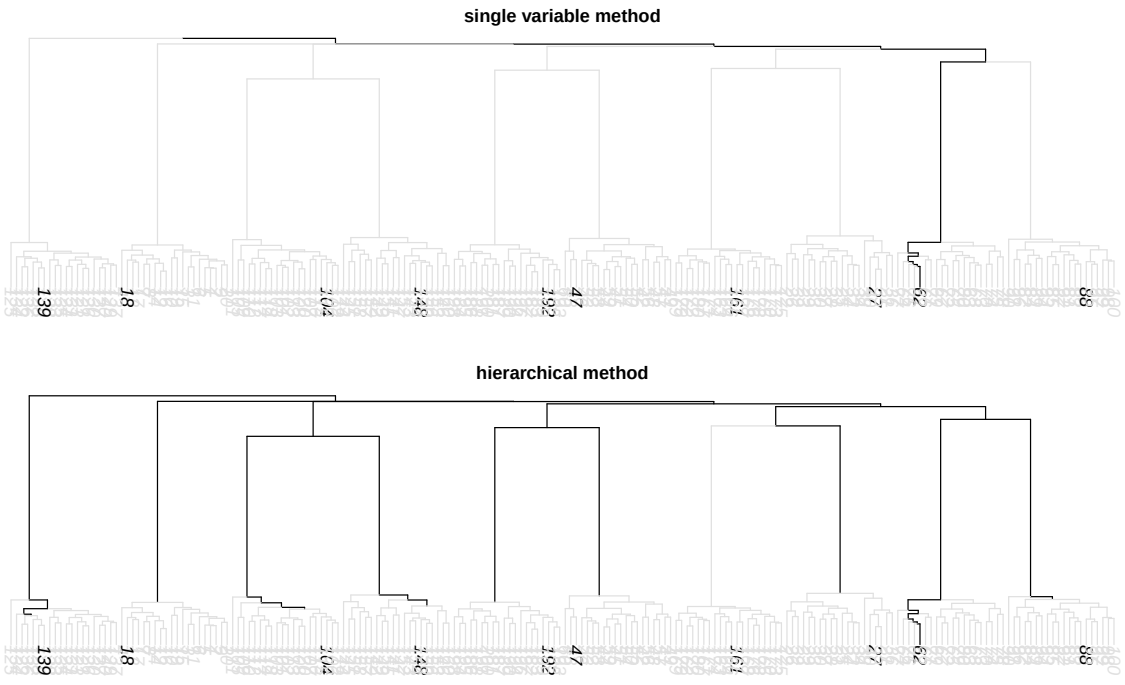


α-weight distribution with inheritance procedure (Goeman and Finos, 2012)



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



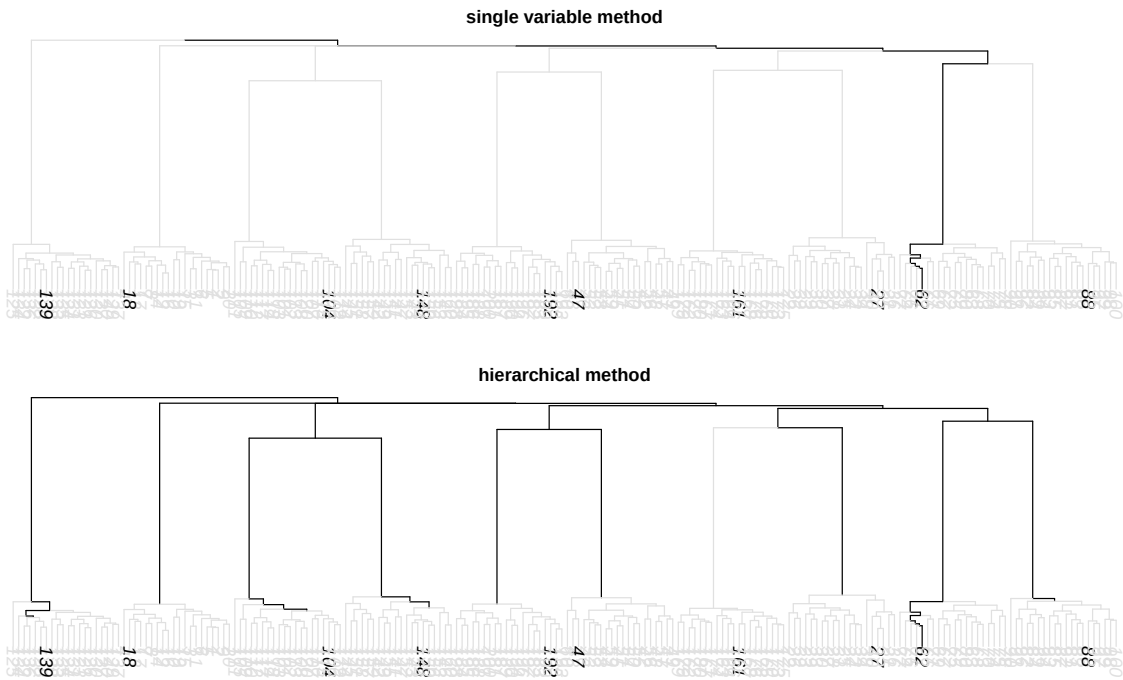
and avoid to test all possible subset of groups...!!!
which would be a disaster from a computational and multiple testing adjustment point of view

Does this work?

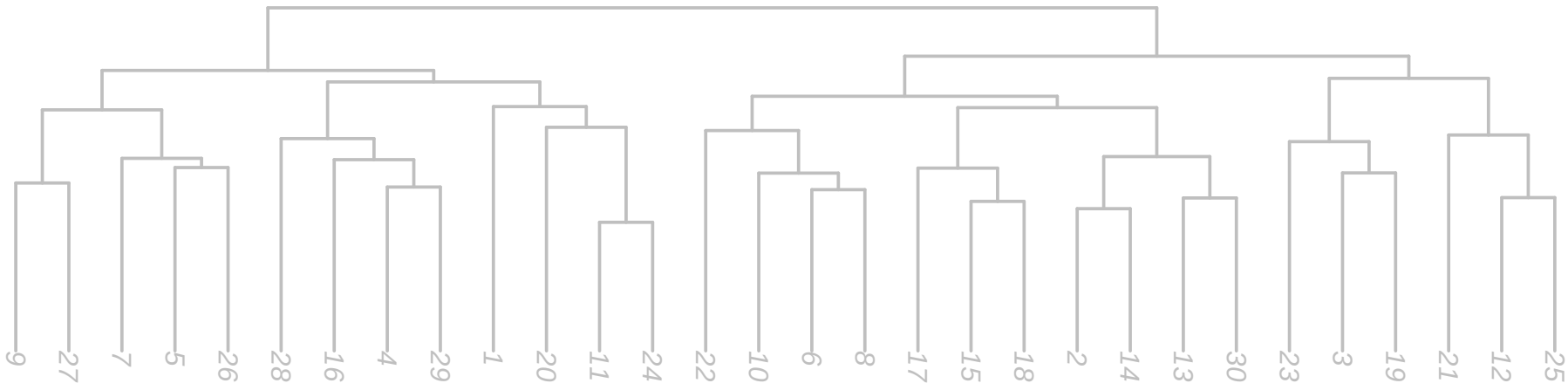
Mandozzi and PB (2014, 2015) provide some theory, implementation and empirical results for simulation study

when using the multiple sample splitting method
(using the desparsified Lasso is more straightforward)

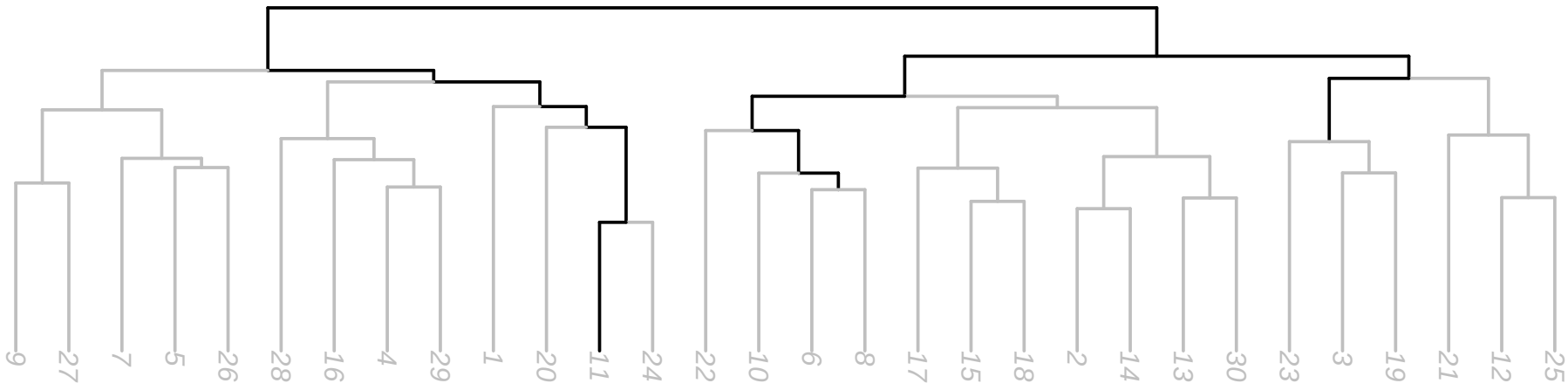
- fairly reliable type I error control
- reasonable power (and clearly better than single variable testing method)



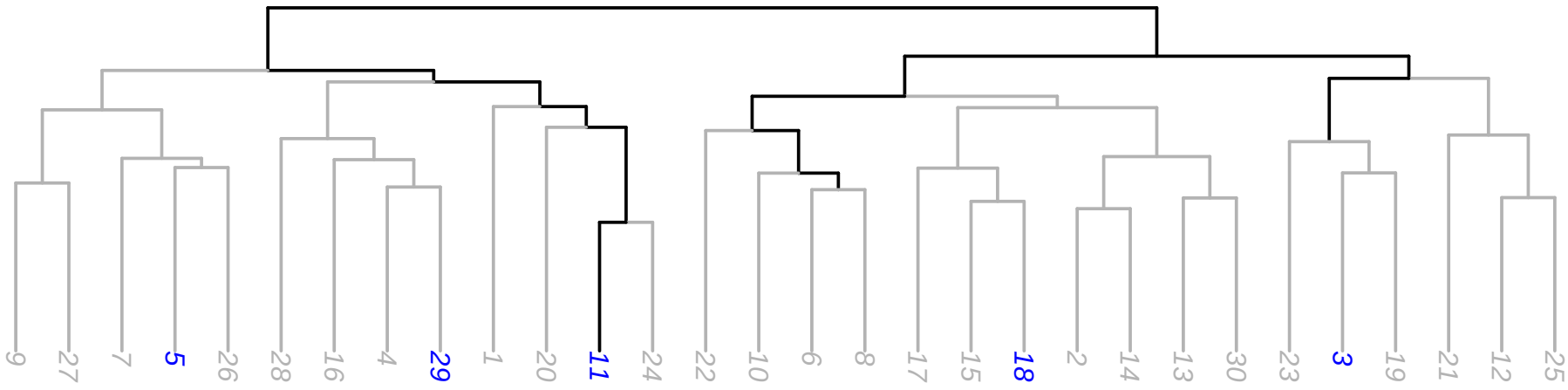
an illustration



an illustration

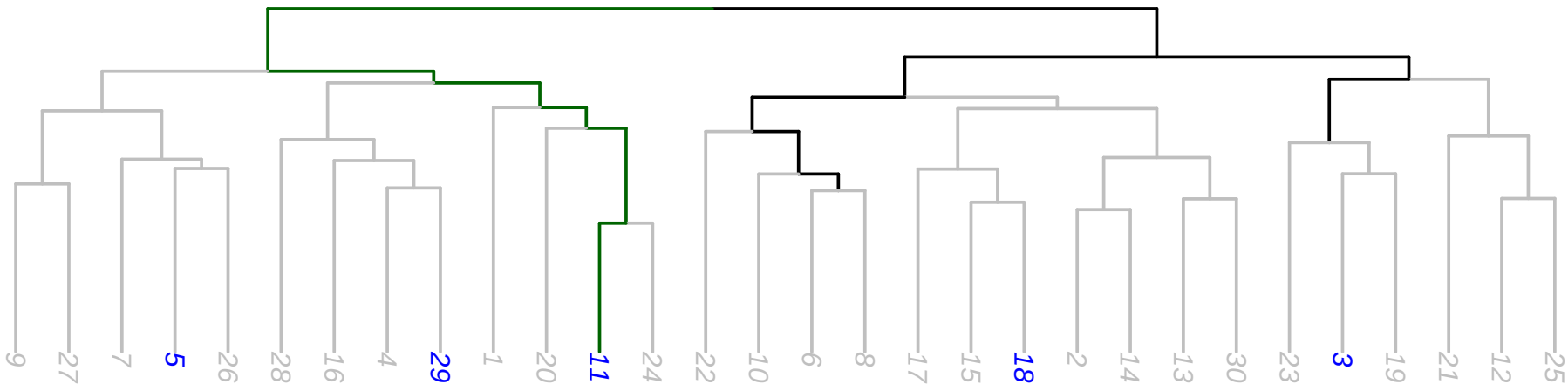


an illustration



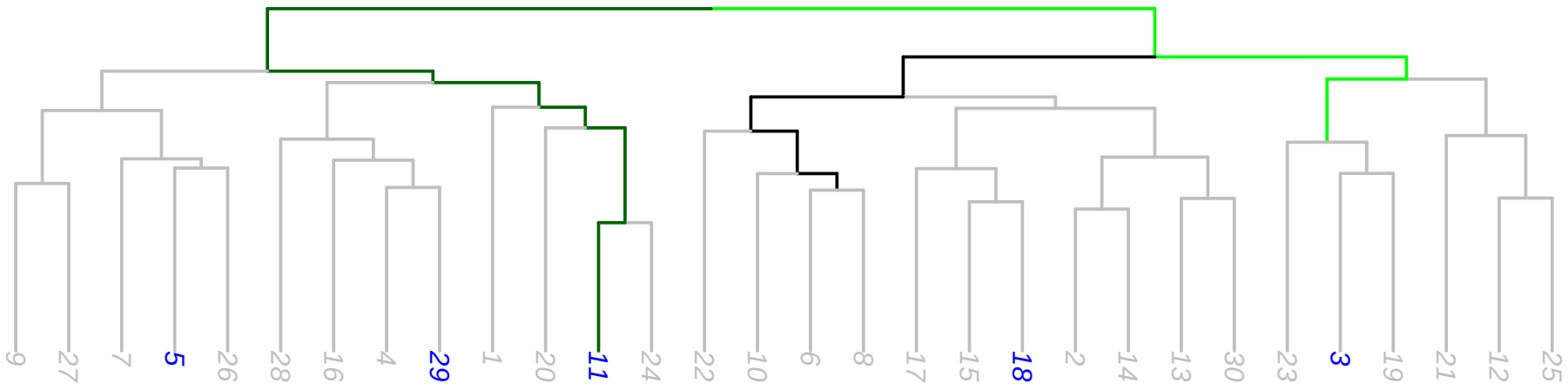
$S_0 = \{5, 29, 11, 18, 3\}$

an illustration



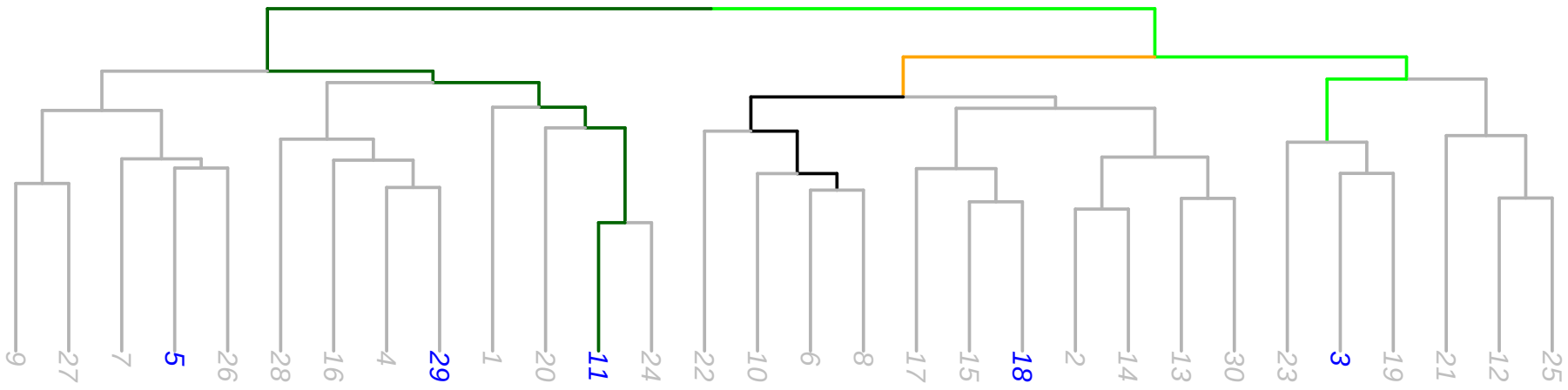
$S_0 = \{5, 29, 11, 18, 3\}$, one STD: $\{11\}$

an illustration



$S_0 = \{5, 29, 11, 18, 3\}$, one STD: $\{11\}$,
one GTD of cardinality 3: $\{23, 3, 19\}$

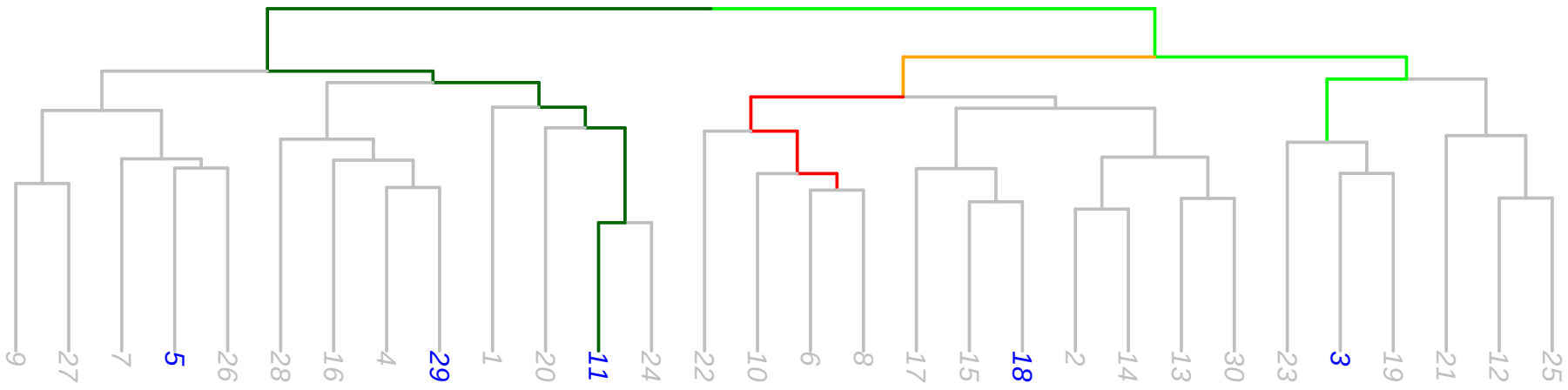
an illustration



$S_0 = \{5, 29, 11, 18, 3\}$, one STD: $\{11\}$,
one GTD of cardinality 3: $\{23, 3, 19\}$

still OK, potential GTD

an illustration



$S_0 = \{5, 29, 11, 18, 3\}$, one STD: $\{11\}$,
one GTD of cardinality 3: $\{23, 3, 19\}$

still OK, potential GTD , false detection!

A “real” test: GWAS (Buzdugan, Kalisch, Schunk, Fehr and PB, 201x)

motivation: find significant associations in the behavioral economy data

next step: validate the hierarchical inference methodology on a much better studied problem
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. 380’000 SNPs per individuum

Crohn's disease

small groups

SNP group size	chrom.	gene	p-value	hit
7	1	IL23R	0.018	yes
1	2	ATG16L1	$7 \cdot 10^{-6}$	yes
44	5	intergenic	0.009	yes
6	10	LINC01475	0.042	yes
3	10	ZNF365	0.030	yes
1	16	NOD2	$2 \cdot 10^{-4}$	yes
1	18	intergenic	0.040	yes

some single SNPs are found as significant!

"hit": SNP (in the group) is found by WTCCC or by WTCCC replication studies

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

Bipolar disease

only large groups/clusters are found as significant
$\rightsquigarrow$ that's "OK"...

Behavioral economics and genomewide association

with Ernst Fehr, University of Zurich

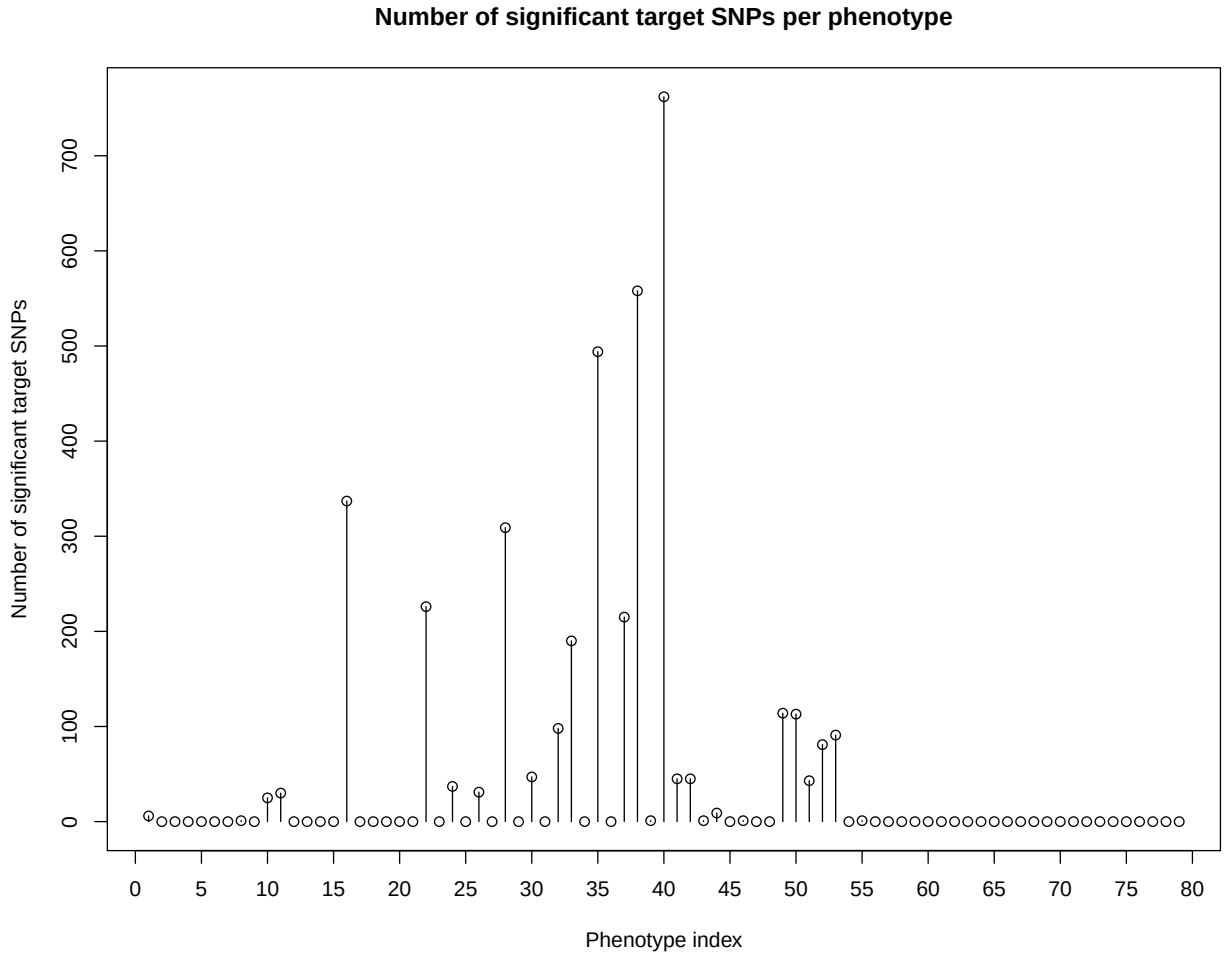
- $n = 1525$ probands (all students!)
- $m = 79$ response variables measuring various behavioral characteristics (e.g. risk aversion) from well-designed experiments
- $p \approx 0.5 \cdot 10^6$ SNPs (the same SNPs per response)

model: multivariate linear model

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

$\rightsquigarrow$ perform hierarchical inference (of course...)

number of significant SNP parameters per response



response 40 has most significant groups of SNPs

I cannot tell more at the moment...

Software

`R`-package `hdi` (Meier, 2013)

contains

- de-sparsified Lasso, Ridge projection method, multiple sample splitting, stability selection
- hierarchical inference

Conclusions

key concepts for high-dimensional statistics:

- sparsity of the underlying regression vector
  - sparse estimator is optimal for prediction/estimation
  - non-sparse estimators are optimal for uncertainty quantification

due to near collinearity of a few covariables (which is to be expected with $p \gg n$)
$\rightsquigarrow$ inference for single variables is often ill-posed

hierarchical inference is a good way to address these issues

in view of (yet) uncheckable assumptions

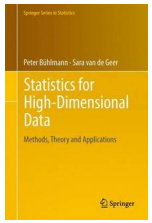
$\rightsquigarrow$

confirmatory high-dimensional inference
remains an interesting challenge

Thank you!

References:

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



- Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. Journal of the American Statistical Association 104, 1671-1681.
- Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. Bernoulli 19, 1212-1242.
- van de Geer, S., Bühlmann, P. and Ritov, Y. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics 42, 1166-1202.
- Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2014). High-dimensional inference: confidence intervals, p-values and R-software hdi. Preprint arXiv:1408.4026