High-Dimensional Statistics

Peter Bühlmann
ETH Zürich

main collaborators

Sara van de Geer

Nicolai Meinshausen

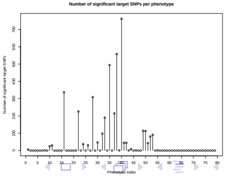
High-dimensional data

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."

statistics is important...

and its mathematical roots as well !

statistics is important...

and its mathematical roots as well !

Linear model

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

standard vector- and matrix-notation:

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

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

- design matrix X: either deterministic or stochastic
- error/noise ε:
  $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d., $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$
  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: measures the effect which is not explained by the other covariables for stochastic $X = (X^{(1)}, \ldots, X^{(p)})^T$ with $\mathrm{Cov}(X) = \Sigma_{p \times p}$:

$$\beta^0 = \Sigma^{-1} \begin{pmatrix} \mathrm{Cov}(Y, X^{(1)}) \\ \cdots \\ \cdots \\ \mathrm{Cov}(Y, X^{(p)}) \end{pmatrix}$$

complicated expression with Σ^{-1}! particularly if p is large

note that β^0_j depends on whether there are many or only a few other covariables $\{X_k;\ k \neq j\}$

in contrast: marginal correlation

$$\rho_{Y,j} = \mathrm{Cor}(Y, X^{(j)})$$

remains the same regardless whether there are no or many other variables $\{X^{(k)};\ k \neq j\}$!

why making it complicated... ?

because

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

is often the much more appropriate quantity in applications

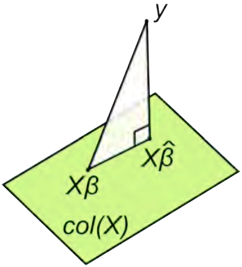
we want to measure the effect of $X^{(j)}$ on Y which has not been explained by the other covariables $\{X^{(k)};\ k \neq j\}$

Least squares solution

based on data $Y_{n\times 1}$, $X_{n\times p}$:
want to estimate the unknown regression parameter β^0

(ordinary) least squares:

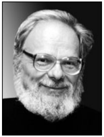
$$\hat{\beta}_{\mathrm{LS}} = \mathrm{argmin}_{\beta}\|Y - X\beta\|_2^2,$$
$$\hat{\beta}_{\mathrm{LS}} = (X^T X)^{-1} X^T Y$$



cannot be used...

we could use generalized least squares... but the minimizer is not unique and residual sum of squares equals zero
$\leadsto$ statistical overfitting!
the estimate would be very poor for prediction on new data

Regularization



ℓ_2-norm regularization (Tikhonov 1943, 1963)
or Ridge regression (Hoerl, 1962; Hoerl and Kennard, 1970)

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \text{argmin}_{\beta}(\|Y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2),$$

- unique and explicit solution:

$$\hat{\beta}_{\ell_2-\text{regul.}} = (X^T X/n + \lambda I)^{-1} X^T Y/n$$

but...

- poor prediction power (if truth is sparse and "non-smooth")
  not a sparse solution: impractical, no easy interpretation

ℓ_0-regularization

$$\hat{\beta}_{\ell_0-\text{regul.}} = \text{argmin}_\beta(\|Y - X\beta\|_2^2/n + \lambda \underbrace{\|\beta\|_0^0}_{\text{no. of non-zero comp.}})$$

AIC (Akaike, 1970),... , BIC (Schwarz, 1978),...

- solution is typically unique and sparse
  but ...
- impossible to compute (NP hard in general)

ℓ_1-norm regularization

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

also called Lasso (Tibshirani, 1996):

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

convex optimization problem

- sparse solution (because of "ℓ_1-geometry")
- not unique in general... but unique with high probability under some assumptions (see later)

LASSO = Least Absolute Shrinkage and Selection Operator

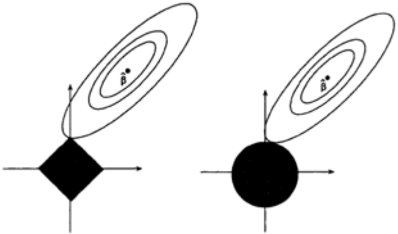
more about "ℓ_1-geometry"

equivalence to primal problem

$$\hat{\beta}_{\mathrm{primal}}(R) = \mathrm{argmin}_{\beta;\|\beta\|_1 \le R} \|Y - X\beta\|_2^2/n,$$

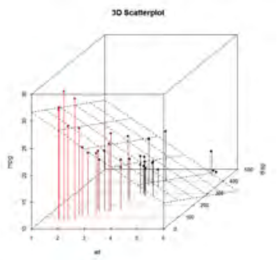
with a one-to-one correspondence between λ and R which depends on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$

p=2



left: ℓ_1-"world"
residual sum of squares reaches a minimal value (for certain constellations of the data) if its contour lines hit the ℓ_1-ball in its corner $\rightsquigarrow$ $\hat{\beta}_1 = 0$

Prediction and estimation of the regression surface



predict new (future) response variables Y_{new} with corresponding design matrix X

$$\mathbb{E}_{Y_{\text{new}}} \| Y_{\text{new}} - X\hat{\beta}\|_2^2/n = \underbrace{\|X(\hat{\beta} - \beta^0)\|_2^2/n}_{\text{error for true regression surface}} + \underbrace{\sigma^2}_{=\text{const.}}$$

question: under which assumptions can we achieve

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = o_P(1) \ (p \geq n \to \infty)$$

under which assumptions can we achieve

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = o_P(1) \; (p \geq n \to \infty)$$

note: for least squares estimator:

$$\|X(\hat{\beta}_{\mathrm{LS}} - \beta^0)\|_2^2/n = \|Y - X\beta^0\|_2^2/n \asymp \sigma^2 \neq o_P(1)!$$

because of overfitting

and the same is true for Ridge estimation (ℓ_2-norm regularization) under which assumptions can we achieve

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = o_P(1) \ (p \geq n \to \infty)$$

note: for least squares estimator:

$$\|X(\hat{\beta}_{\mathrm{LS}} - \beta^0)\|_2^2/n = \|Y - X\beta^0\|_2^2/n \asymp \sigma^2 \neq o_P(1)!$$

because of overfitting

and the same is true for Ridge estimation (ℓ_2-norm regularization)

Analysis of Lasso (ℓ_1-norm regularization)

Basic inequality

$$n^{-1}\|X(\hat{\beta}-\beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq 2n^{-1}\varepsilon^T X(\hat{\beta}-\beta^0) + \lambda\|\beta^0\|_1$$

Proof:

$$n^{-1}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq n^{-1}\|Y - X\beta^0\|_2^2 + \lambda\|\beta^0\|_1$$

$$n^{-1}\|Y - X\hat{\beta}\|_2^2 = n^{-1}\|X(\hat{\beta}-\beta^0)\|_2^2 + n^{-1}\|\varepsilon\|_2^2 - 2n^{-1}\varepsilon^T X(\hat{\beta}-\beta^0)$$
$$n^{-1}\|Y - X\beta^0\|_2^2 = n^{-1}\|\varepsilon\|_2^2$$

$\leadsto$ statement above □

need a bound for $2n^{-1}\varepsilon^T X(\hat{\beta}-\beta^0)$

$$2n^{-1}\varepsilon^T X(\hat{\beta} - \beta^0) \ \ \le \ \ 2 \max_{j=1,\ldots,p} |n^{-1} \sum_{i=1}^{n} \varepsilon_i X_i^{(j)}| \|\hat{\beta} - \beta^0\|_1$$

consider

$$\mathcal{F}(\lambda_0) = \{2 \max_{j} |n^{-1} \sum_{i=1}^{n} \varepsilon_i X_i^{(j)}| \le \lambda_0\}$$

the probabilistic part of the problem

$$\text{on } \mathcal{F}(\lambda_0): \quad 2n^{-1}\varepsilon^T X(\hat{\beta} - \beta^0) \ \ \le \ \ \lambda_0 \|\hat{\beta} - \beta^0\|_1 \le \lambda_0 \|\hat{\beta}\|_1 + \lambda_0 \|\beta^0\|_1$$

and hence using the Basic inequality

$$\text{on } \mathcal{F}(\lambda_0): \quad n^{-1}\|X(\hat{\beta} - \beta^0)\|_2^2 + (\lambda - \lambda_0)\|\hat{\beta}\|_1 \le (\lambda_0 + \lambda)\|\beta^0\|_1$$

for $\lambda \ge 2\lambda_0$:

$$\text{on } \mathcal{F}(\lambda_0) = \mathcal{F}(\lambda_0): \quad 2n^{-1}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \le 3\lambda\|\beta^0\|_1$$

Consistency of Lasso (under weak conditions)

Theorem (Greenshtein & Ritov, 2004; PB & van de Geer, 2011)
On the set

$$\mathcal{F} = \{4 \max_{j=1,\ldots,p} |\varepsilon^T X^{(j)}/n| \le \lambda\}:$$

$$\|X(\hat{\beta}(\lambda) - \beta^0)\|_2^2/n \le \frac{3}{2}\lambda\|\beta^0\|_1$$

$\rightsquigarrow$ trade-off for choosing λ:

- small λ: good accuracy but with low probability
- large λ: poor accuracy with high probability

if $\|\beta^0\|_1 = o(\lambda^{-1}) \underbrace{=}_{\lambda \asymp \sqrt{\log(p)/n}} o(\sqrt{n/\log(p)})$ "OK" if $\log(p) \ll n$

$\Longrightarrow$ convergence to zero

Consistency of Lasso (under weak conditions)

Theorem (Greenshtein & Ritov, 2004; PB & van de Geer, 2011)
On the set

$$\mathcal{F} = \{4 \max_{j=1,\ldots,p} |\varepsilon^T X^{(j)}/n| \le \lambda\} :$$

$$\|X(\hat{\beta}(\lambda) - \beta^0)\|_2^2/n \le \frac{3}{2}\lambda\|\beta^0\|_1$$

$\rightsquigarrow$ trade-off for choosing λ:

- small λ: good accuracy but with low probability
- large λ: poor accuracy with high probability

if $\|\beta^0\|_1 = o(\lambda^{-1}) \underbrace{=}_{\lambda \asymp \sqrt{\log(p)/n}} o(\sqrt{n/\log(p)})$ "OK" if $\log(p) \ll n$

$\Longrightarrow$ convergence to zero

recap: the proof is based on decoupling into

- a deterministic part (easy to derive)
- a probabilistic part (the set $\mathcal{F}$)

Probability of $\mathcal{F}$ and choice of λ

$$\text{if } \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I) \Longrightarrow \varepsilon^T X^{(j)}/n \sim \mathcal{N}(0, \underbrace{\|X^{(j)}\|_2^2/n}_{\text{standardized}=1} \cdot \frac{1}{n})$$

$\leadsto$

$$\mathbb{P}[\max_{j=1,\ldots,p} |\varepsilon^T X^{(j)}/n| > c] \leq 2p \exp(-c^2 n/(2\sigma^2))$$

$\leadsto$ for $\lambda = 4\sigma\sqrt{\frac{t^2+2\log(p)}{n}}$

$$\mathbb{P}[\mathcal{F}] \geq 1 - 2\exp(-t^2/2)$$

in short: $\lambda \asymp \sqrt{\log(p)/n}$ leads to $\mathbb{P}[\mathcal{F}] \approx 1$

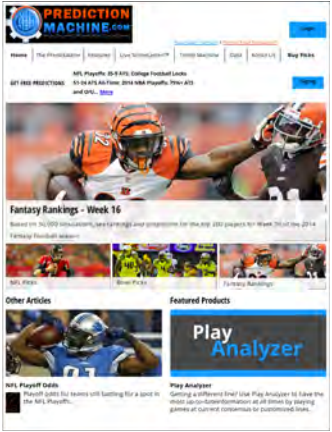
Corollary

assume Gaussian errors

for $\lambda \asymp \sqrt{\log(p)/n}$: $\|X(\hat{\beta}(\lambda) - \beta^0)\|_2^2/n = O_P(\sqrt{\log(p)/n}\|\beta^0\|_1)$

Lasso is a popular machine for prediction in numerous applications

Sports Picks: Free NFL, MLB, College Basketball... http://www.predictionmachine.com/



1 von 10 12/18/2014 11:04 AM

computational biology/bioinformatics, climate research, economics/econometrics, imaging, ...

can easily generalize to
non-Gaussian errors, dependent errors,...

need to control

$$\mathbb{P}[\max_j |\varepsilon^T X^{(j)}/n| > c]$$

Example: $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d., $\mathbb{E}|\varepsilon_i|^2 \le C_1 < \infty$,
$\max_j \|X_i^{(j)}\|_\infty \le C_2 < \infty$
use Nemirovski's inequality: for $Z_1, \ldots, Z_n$ independent,

$$\mathbb{E}[\max_j |\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])|^m] \le (8\log(2p))^{m/2}\mathbb{E}[\max_j \sum_{i=1}^n Z_i^2]^{m/2}$$

$$\Longrightarrow \max_j |\varepsilon^T X^{(j)}/n| = O_P(\sqrt{\log(p)/n})$$

Estimation of parameters ("inverse problem")

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

with fixed (deterministic) design X

goal: inferring the unknown β^0 (instead of $X\beta^0$)

problem of identifiability:
for $p > n$: $X\beta^0 = X\theta$
for any $\theta = \beta^0 + \xi$, ξ in the null-space of X

$\leadsto$ cannot identify β^0 without further assumptions!
(in contrast to prediction...)

Compressed sensing (in the noiseless case)

(Candes & Tao, 2005; Donoho& Huo, 2001; ...)

linear measurements $Y = X\beta^0$ with X known

goal: recover p-dimensional β^0 (e.g. the unknown pixel-intensities of an image) from under-sampled measurements Y

ℓ_1-problem:

$$\hat{\beta} = \mathrm{argmin}_{\beta} \|\beta\|_1 \text{ such that } Y = X\beta$$

assume

- β^0 is ℓ_0-sparse (having s_0 non-zero coefficients)
- X is "sufficiently nice" (restricted isometry)
  for $n < p$: probabilistic results that restricted isometry holds

$\leadsto$ exact recovery $\hat{\beta} = \beta^0$

many generalizations to noisy case

$\rightsquigarrow$ equivalence to the problem from high-dimensional statistics

Restricted eigenvalues (for identifiability)

suppose $X\theta = X\beta^0$

$$0 = \|X(\theta - \beta^0)\|_2^2/n = (\theta - \beta^0)^T \underbrace{\hat{\Sigma}}_{X^T X/n} (\theta - \beta^0)$$

$\leadsto$ if $\hat{\Sigma}$ were invertible $\Longrightarrow \theta = \beta^0$

"quantify" ill-posedness with minimal eigenvalue $\Lambda^2_{\min}(\hat{\Sigma})$ of $\hat{\Sigma}$:

$$\forall \beta: \ \|\beta\|_2^2 \le \frac{\beta^T \hat{\Sigma} \beta}{\Lambda^2_{\min}(\hat{\Sigma})}$$

with $p > n$: $\Lambda^2_{\min}(\hat{\Sigma}) = 0$ ...

smallest restricted ℓ_1-eigenvalue (van de Geer, 2007)

active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$ with $s_0 = |S_0|$

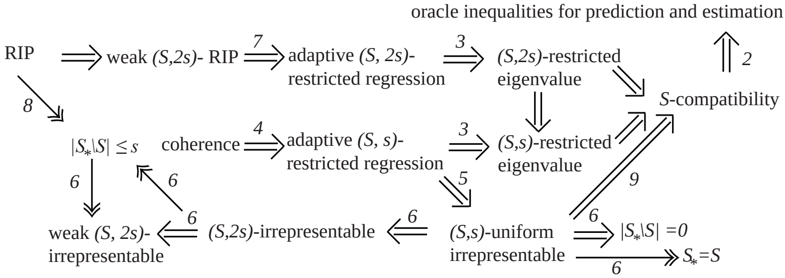
smallest restricted eigenvalue $\phi_0^2 > 0$:

for all β satisfying $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$

$$\|\beta_{S_0}\|_1^2 \leq \frac{(\beta^T \hat{\Sigma} \beta) s_0}{\phi_0^2}$$

(appearance of s_0 due to $\|\beta_{S_0}\|_1^2 \leq s_0 \|\beta_{S_0}\|_2^2$)

various conditions and their relations (van de Geer & PB, 2009)



smallest restricted eigenval. is (substantially) weaker than RIP

Theorem (PB & van de Geer, 2011)

- X has i.i.d. rows with sub-Gaussian distribution
- $\mathrm{Cov}(X_i) = \Sigma$ has smallest eigenvalue $\Lambda^2_{\min}(\Sigma) \geq C > 0$
  e.g. Σ is Toeplitz matrix; or equi-corr. with $0 < \rho < 1$

if $s_0 =$ no. of non-zero coefficients in $\beta^0 = o(\sqrt{n/\log(p)})$,
with high probability:

smallest restricted ℓ_1-eigenvalue of $\hat{\Sigma}$ satisfies: $\phi_0^2 > C/2$

consider Lasso

$$\hat{\beta}(\lambda) = \mathrm{argmin}_{\beta}(n^{-1}\|Y - X\beta\|^2 + \lambda\|\beta\|_1)$$

assuming restricted ℓ_1-eigenvalue (compatibility) condition:
for $\lambda \asymp \sqrt{\log(p)/n}$:

$$n^{-1}\|X(\hat{\beta} - \beta^0)\|_2^2 \le O_P(s_0 \log(p)/n)$$
$$\|\hat{\beta} - \beta^0\|_1 \le O_P(s_0\sqrt{\log(p)/n})$$

$s_0 = |S_0|$ is the cardinality of the active set

that is:

$$\beta^0 \text{ is identifiable if } \underbrace{s_0 \ll \sqrt{n/\log(p)}}_{\text{sparse !}}$$

"sketch" of proof:
recall the basic inequality

$$n^{-1}\|X(\hat{\beta}-\beta^0)\|_2^2+\lambda\|\hat{\beta}\|_1 \le 2n^{-1}\varepsilon^T X(\hat{\beta}-\beta^0)+\lambda\|\beta^0\|_1$$

simple re-writing (triangle inequality) on $\mathcal{F}(\lambda)$,

$$2\|(\hat{\beta}-\beta^0)\hat{\Sigma}(\hat{\beta}-\beta^0)\|_2^2+\lambda\|\hat{\beta}_{S_0^c}\|_1 \le 3\lambda\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1$$

where $\hat{\Sigma}=n^{-1}X^TX$

relate $\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1$ to (with $\le$ relation) $(\hat{\beta}-\beta^0)\hat{\Sigma}(\hat{\beta}-\beta^0)$

$\rightsquigarrow$ invoke (compatibility) restricted ℓ_1-eigenvalue condition

$\rightsquigarrow$ oracle inequality

$$\|X(\hat{\beta}-\beta^0\|_2^2/n+\lambda\|\hat{\beta}-\beta^0\|_1 \le 4\lambda^2 s_0/\phi_0^2$$

$\square$

Lasso-workhorse: Variable screening assuming beta-min condition

$$S_0 = \{j;\ \beta_j^0 \neq 0\}, \quad \hat{S} = \{j;\ \hat{\beta}_j \neq 0\}$$

(asking for $\hat{S} = S_0$ is often too ambitious)

- "beta-min" condition:

$$\min_{j \in S_0} |\beta_j^0| \gg s_0 \sqrt{\log(p)/n} \quad \left(\text{or } \sqrt{s_0 \log(p)/n} \text{ or } \sqrt{\log(p)/n}\right)$$

- (compatibility) restricted ℓ_1-eigenv. condition:

from $\|\hat{\beta} - \beta^0\|_1 \leq O_P(s_0 \sqrt{\log(p)/n})$ we immediately obtain

variable screening: $\hat{S} \supseteq S_0$ with high probability

and: $|\hat{S}| \leq \min(n, p)$

i.e., we will not miss a true variable!

but we may (typically) have too many false positive selections

Lasso-workhorse: Variable screening assuming beta-min condition

$$S_0 = \{j;\ \beta_j^0 \neq 0\}, \quad \hat{S} = \{j;\ \hat{\beta}_j \neq 0\}$$

(asking for $\hat{S} = S_0$ is often too ambitious)

- "beta-min" condition:

$$\min_{j \in S_0} |\beta_j^0| \gg s_0 \sqrt{\log(p)/n} \quad \left(\text{or } \sqrt{s_0 \log(p)/n} \text{ or } \sqrt{\log(p)/n}\right)$$

- (compatibility) restricted ℓ_1-eigenv. condition:

from $\|\hat{\beta} - \beta^0\|_1 \leq O_P(s_0\sqrt{\log(p)/n})$ we immediately obtain

variable screening: $\hat{S} \supseteq S_0$ with high probability

and: $|\hat{S}| \leq \min(n, p)$

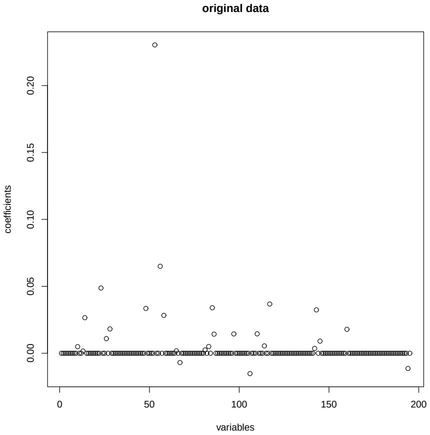
i.e., we will not miss a true variable!

but we may (typically) have too many false positive selections

Example: motif regression (computational biology)

$p = 195, n = 143$

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



which variables in $\hat{S}$ are false positives?
p-values/quantifying uncertainty would be very useful!

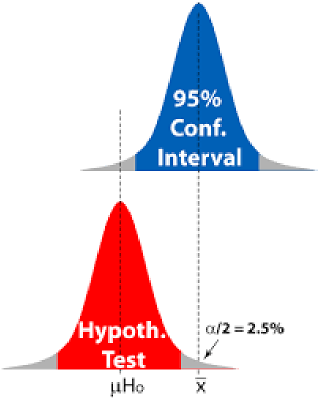
remember the conditions for $\hat{S} \supseteq S_0$:

- (compatibility) restricted ℓ_1-eigenv. condition for X
  $\leadsto$ "unavoidable"
- beta-min condition (strong assumption!)
  and we will relax this in the sequel

remember the conditions for $\hat{S} \supseteq S_0$:

- (compatibility) restricted ℓ_1-eigenv. condition for X
  $\leadsto$ "unavoidable"
- beta-min condition (strong assumption!)
  and we will relax this in the sequel

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
- ...

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

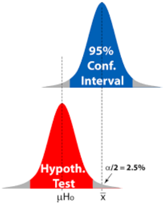
classical goal: statistical hypothesis testing

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

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

$\leadsto$ 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:

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

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

Asymptotic pivot and optimality

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

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

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 !

Asymptotic pivot and optimality

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

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

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

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

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

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

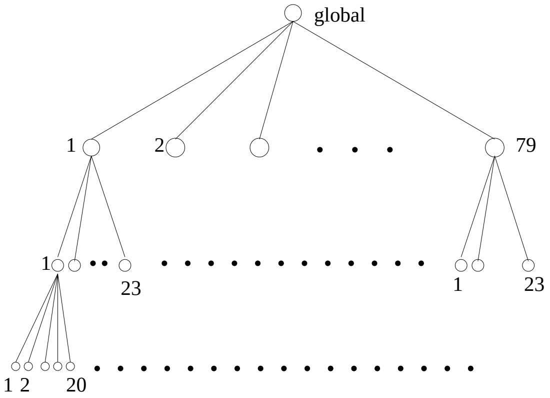
interested in p-values for

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

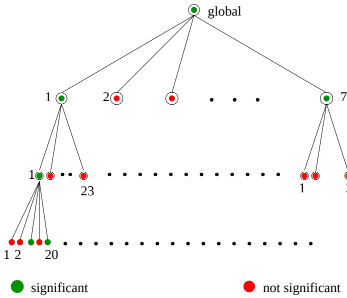
adjusted for multiple testing (among very many hypotheses!)

there is structure!

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



do hierarchical FWER adjustment (Meinshausen, 2008)

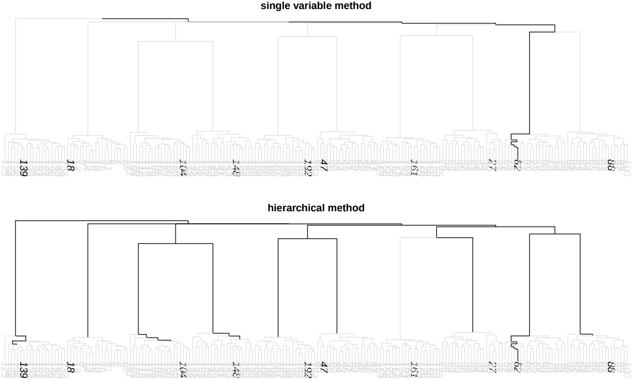


1. test global hypothesis
2. if significant: test all single response hypotheses
3. for the significant responses: test all single chromosome hyp.
4. for the significant chromosomes: test all groups of SNPs

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

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

Mandozzi & PB (2013, 2015):



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

Meinshausen (2008), 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

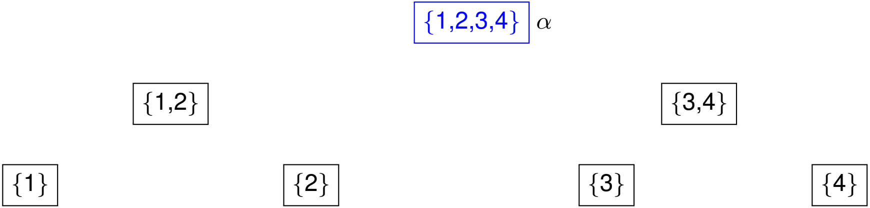
(Meinshausen, 2008)

$$\text{reject } H_{0,G} \text{ if } P_{G;\mathrm{hier-adj}} \leq \alpha$$

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

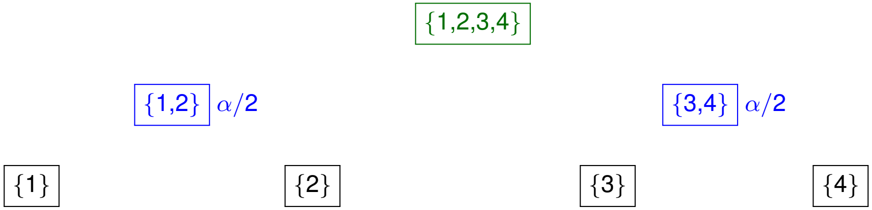
optimizing the procedure:
α-weight distribution with inheritance (Goeman and Finos, 2012)

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



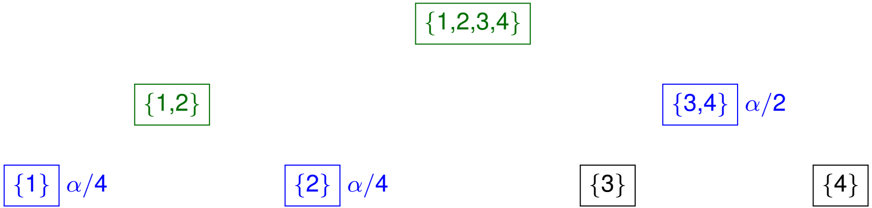
α-weight distribution with inheritance procedure

(Goeman and Finos, 2012)

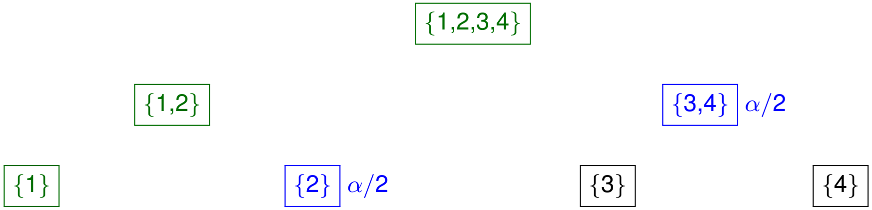


α-weight distribution with inheritance procedure

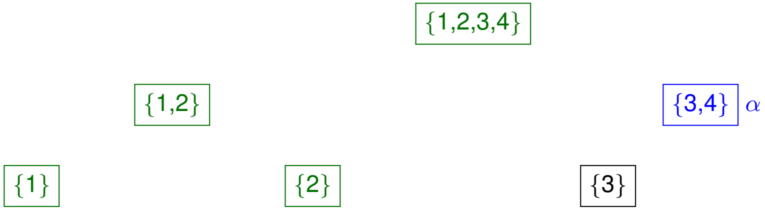
(Goeman and Finos, 2012)



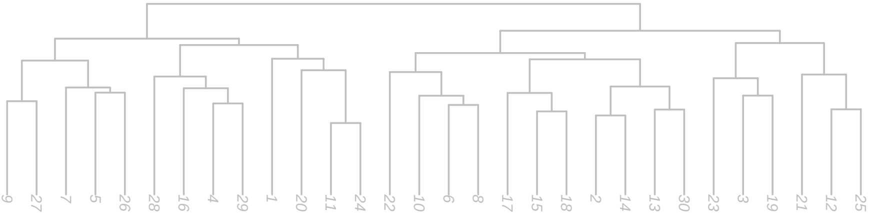
α-weight distribution with inheritance procedure (Goeman and Finos, 2012)



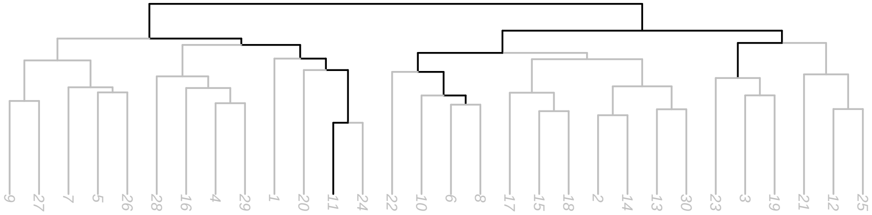
α-weight distribution with inheritance procedure
(Goeman and Finos, 2012)



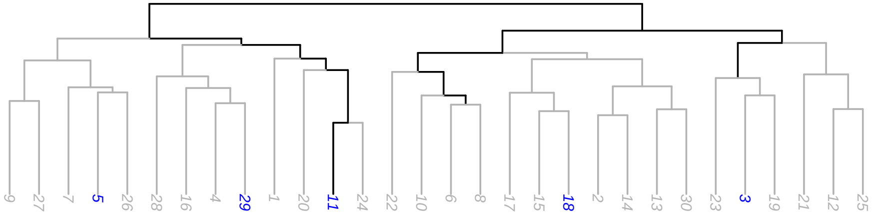
another illustration



another illustration

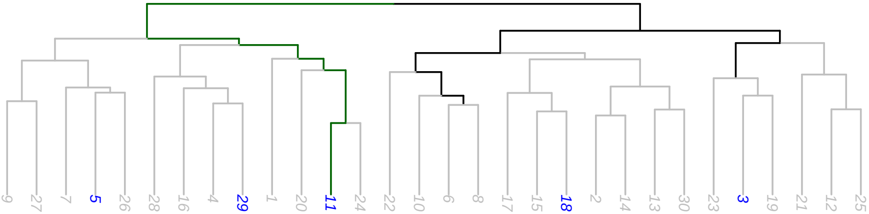


another illustration



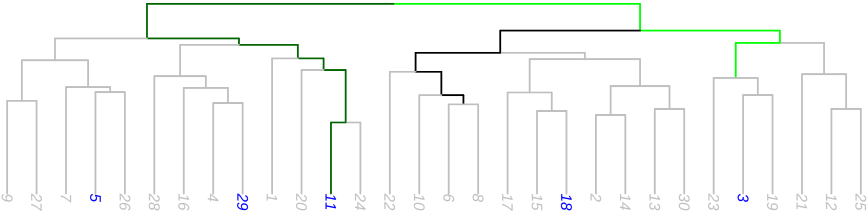
$S_0 = \{5, 29, 11, 18, 3\}$

another illustration



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

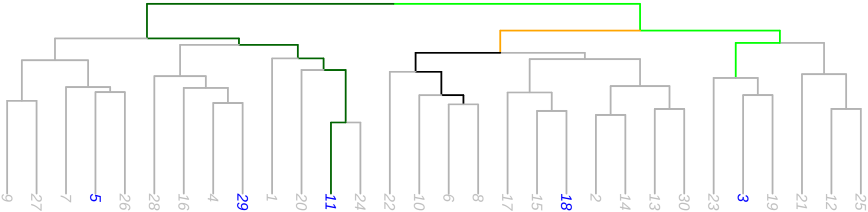
another illustration



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

one GTD of cardinality 3: $\{23, 3, 19\}$

another illustration

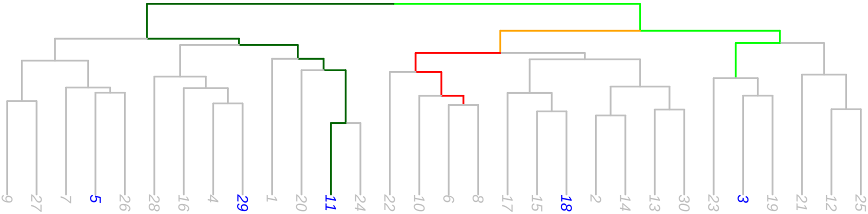


$S_0 = \{5, 29, 11, 18, 3\}$,   one STD: $\{11\}$,

one GTD of cardinality 3: $\{23, 3, 19\}$

still OK, potential GTD

another illustration

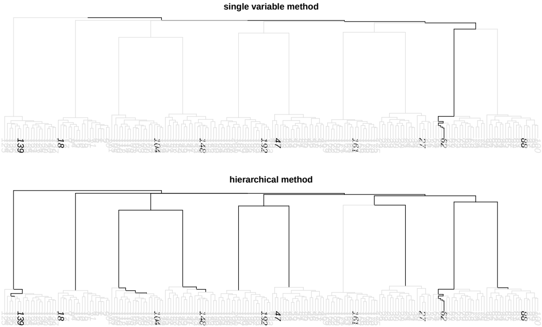


$S_0 = \{5, 29, 11, 18, 3\}$, one STD: $\{11\}$,
one GTD of cardinality 3: $\{23, 3, 19\}$

still OK, potential GTD , false detection!

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

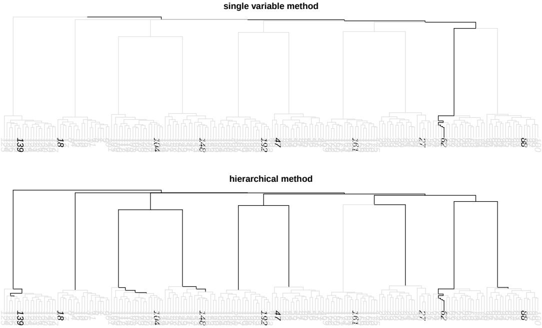


and avoid to test all possible subset of groups...!!!
which would be a disaster from a computational and multiple testing adjustment point of view

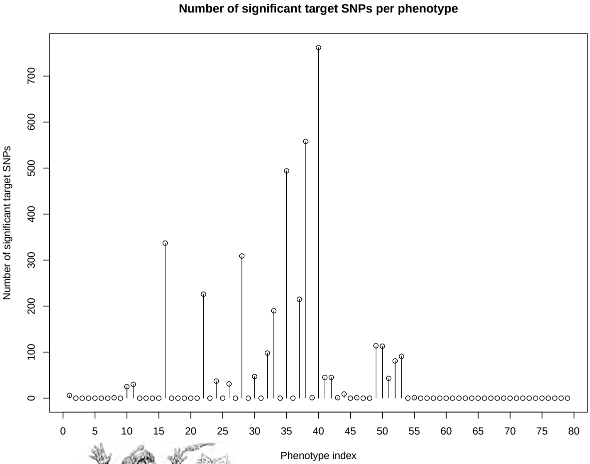
Does this work?

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

- 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

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	$1.7 \cdot 10^{-[illegible]}$	0.013
CD	rs11805303, rs2201841, rs11209033, rs12141431, rs12139179	1	IL23R	$4.5 \cdot 10^{-4}$	0.014
CD	rs10210302	2	ATG16L1	$4.6 \cdot 10^{-[illegible]}$	0.014
CD	rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934	5	intergenic	$2.7 \cdot 10^{-[illegible]}$	0.016
CD	rs10883371	10	LINC01475, NKX2-3	$2.1 \cdot 10^{-2}$	0.004
CD	rs10761659	10	ZNF365	$1.5 \cdot 10^{-[illegible]}$	0.007
CD	rs2076756	16	NOD2	$1.3 \cdot 10^{-[illegible]}$	0.017
CD	rs2542151	18	intergenic	$1.5 \cdot 10^{-[illegible]}$	0.005
RA	rs6679677	1	PHTF1	$5.9 \cdot 10^{-[illegible]}$	0.031
RA	rs9272346	6	HLA-DQA1	$1.4 \cdot 10^{-[illegible]}$	0.017

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

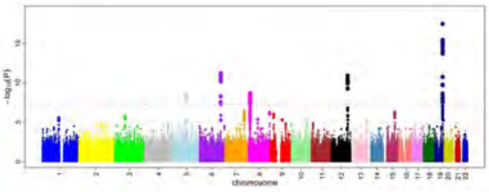
large groups

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



$\leadsto$ significant marginal findings cluster in regions

and then assign ad-hoc regions $+/-10k$ base pairs around the single significant SNPs
still: this is only marginal inference
not the effect of a SNP which is adjusted by the presence of many 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: 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: hierarchical inference is the “first” promising method to infer causal (groups of) SNPs

Genomewide association study in plant biology

Klasen, Barbez, Meier, Meinshausen, PB, Koornneef, Busch & Schneeberger (2015)

root development in Arabidopsis Thaliana

hierarchical inference: 4 significant associations

3 new associations are within and neighboring to PEPR2 gene
$\leadsto$ validation resulted to impact root meristem size

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:

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

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)

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

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

$$\|\beta^0\|_0 \le b_{\max}^2 |S_{f^0}|$$

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

in general: linear projection is less sparse than $f^0(.)$
but ℓ_r-sparsity assump. is sufficient for e.g. de-sparsified Lasso

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

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

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
$\leadsto$ 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

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

we need to adjust the variance formula

(Huber, 1967; Eicker, 1967; White, 1980)

easy to do: e.g. for the de-sparsified Lasso, we compute

$$Z_j = X_j - X_{-j}\hat{\gamma}_j \ \text{Lasso residuals from } X_j \ vs. X_{-j}$$

$$\hat{\varepsilon} = Y - X\hat{\beta} \ \text{Lasso residuals from } Y \ vs. X$$

$$\hat{\omega}_{jj}^2 = \text{empirical variance of } \hat{\varepsilon}_i Z_{j;i} \ (i = 1, \ldots, n)$$

Theorem (PB and van de Geer, 2015)
assume: ℓ_r-sparsity of β^0 $(0 < r < 1)$, $\mathbb{E}|\varepsilon|^{2+\delta} \le K < \infty$,
and ℓ_r-sparsity $(0 < r < 1)$ for rows of $\Sigma = \mathrm{Cov}(X)$:

$$\sqrt{n}\frac{Z_j^T X_j/n}{\hat{\omega}_{jj}}(\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0,1)$$

message:
for random design, inference machinery for projected parameter β^0 "works" when adjusting the variance formula

in addition for Gaussian design:
if a variable is significant in the projected linear model
$\leadsto$ it must be significant in the nonlinear function

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:
it becomes an assumption that there exists β^0 which is sufficiently ℓ_r-sparse ($0 < r \leq 1$)

no new theory is required; adapted variance formula captures heteroscedastic errors

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"

The bootstrap (Efron, 1979): more reliable inference


Efron

residual bootstrap for fixed design:

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

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

i.i.d. resampling of centered residuals $\rightsquigarrow \varepsilon_1^*, \ldots, \varepsilon_n^*$

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

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

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

compute algorithm/estimator $g(\cdot)$ on $\{(X_i, Y_i^*)\}_{i=1}^n$ many times to approximate the true distribution of $g(\{X_i, Y_i\}_{i=1}^n)$

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 and X_j vs. X_{-j} real assumption
- errors can be heteroscedastic and non-Gaussian with 4th moments weak assumption

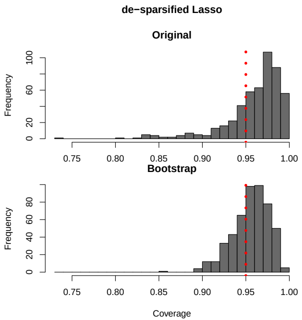
$\leadsto$ consistency of the bootstrap for simultaneous inference!

can approximate

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

(Dezeure, PB & Zhang, 2016)

involves very high-dimensional maxima of non-Gaussian (but limiting Gaussian) quantities (see 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. $p_{\mathrm{eff}} = 600$ instead of $p = 1000$
  or $p_{\mathrm{eff}} = 100K$ 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), in spirit of Westfall and Young (1993):
effective dimension is e.g. $p_{\mathrm{eff}} = 600$ instead of $p = 1000$
or $p_{\mathrm{eff}} = 100K$ instead of $p = 1M$

need to control under the "complete null-hypotheses"

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

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

$\leadsto$ the bootstrap works with (adapts to) effective dimension p_{eff} whereas Bonferroni-Holm adjustment uses "raw" dimension p

Towards model uncertainty

frequentist statistics: goodness of fit of a model

here: null-hypothesis

$$H_0: \; Y = X\beta^0 + \varepsilon \text{ with sparse } \beta^0 \text{ and } \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

alternative: any deviation from H_0

RP (Residual Prediction) test (Shah & PB, 2015)

main idea for $p < n$:

- $PY = X\hat{\beta}_{LS}$ (projection)
- under H_0:

$$R = \frac{(I-P)Y}{\|(I-P)Y\|_2} = \frac{(I-P)\varepsilon}{\|(I-P)\varepsilon\|_2} = \frac{(I-P)Z}{\|(I-P)Z\|_2},\ Z \sim \mathcal{N}(0,1).$$

$\leadsto$ can simulate exactly the scaled residuals via simulation of $\mathcal{N}(0,1)$

- can consider any (measurable) function or algorithm of scaled residuals R:

$$g(R)$$

and compute its distribution exactly under H_0 via simulation of $\mathcal{N}(0,1)$

any (measurable) function of scaled residuals...

example:

scaled residuals $R \overset{\text{nonlinear prediction algorithm}}{\Longrightarrow}$ predicted values $\hat{R}$

$\rightsquigarrow$ residuals $\hat{E} = R - \hat{R} \rightsquigarrow$ test-statistic $T = \|\hat{E}\|_2^2$

- if true model is nonlinear
  $\rightsquigarrow$ signal left in the scaled residuals R from linear model
  $\rightsquigarrow$ T is smaller than if the true model is linear (i.e. H_0)
- exact distribution under H_0 via simulation from $\mathcal{N}(0, 1)$

possible algorithms or functions g:

- detecting potential interactions and nonlinearities:
  $g(\cdot)$ are residual sum of squares (or out of bag estimates for prediction error) when fitting Random Forests to scaled residuals R
- detecting potential heteroscedastic errors:
  $g(\cdot)$ are residual sum of squares (or cross-validation estimate for prediction error) when fitting Lasso to absolute scaled residuals $|R|$
- can test significance of individual variables or groups of variables
- ...

RP tests in high-dimensional problems

least squares residuals are zero $\rightsquigarrow$ no scaled LS-residuals

scaled residuals from Lasso:

$$R = \frac{Y - X\hat{\beta}(\lambda)}{\|Y - X\hat{\beta}(\lambda)\|_2} = \frac{X(\beta^0 - \hat{\beta}(\beta^0, \sigma_\varepsilon Z)) + \sigma_\varepsilon Z}{\|X(\beta^0 - \hat{\beta}(\beta^0, \sigma_\varepsilon Z)) + \sigma_\varepsilon Z\|_2} =: R_\lambda(\beta^0, \sigma_\varepsilon Z),\ Z \sim \mathcal{N}(0, 1)$$

where the second line holds under H_0
idea: simulate the distribution of $R_\lambda(\beta^0, \sigma_\varepsilon Z)$
$\rightsquigarrow$ plug-in estimates

$$\hat{R}_\lambda = R_\lambda(\hat{\beta}_{\mathrm{Lasso}}, \hat{\sigma}_{\varepsilon;\mathrm{Lasso}} Z),\ \ Z \sim \mathcal{N}(0, 1)$$

so that we can simulate via $\mathcal{N}(0, 1)$!

RP tests in high-dimensional problems

least squares residuals are zero $\rightsquigarrow$ no scaled LS-residuals

scaled residuals from Lasso:

$$
\begin{aligned}
R &= \frac{Y - X\hat{\beta}(\lambda)}{\|Y - X\hat{\beta}(\lambda)\|_2} \\
&= \frac{X(\beta^0 - \hat{\beta}(\beta^0, \sigma_\varepsilon Z)) + \sigma_\varepsilon Z}{\|X(\beta^0 - \hat{\beta}(\beta^0, \sigma_\varepsilon Z)) + \sigma_\varepsilon Z\|_2} =: R_\lambda(\beta^0, \sigma_\varepsilon Z), \;\; Z \sim \mathcal{N}(0, 1)
\end{aligned}
$$

where the second line holds under H_0
idea: simulate the distribution of $R_\lambda(\beta^0, \sigma_\varepsilon Z)$
$\rightsquigarrow$ plug-in estimates

$$
\hat{R}_\lambda = R_\lambda(\hat{\beta}_{\text{Lasso}}, \hat{\sigma}_{\varepsilon;\text{Lasso}} Z), \;\; Z \sim \mathcal{N}(0, 1)
$$

so that we can simulate via $\mathcal{N}(0, 1)$!

Theorem (Shah & PB, 2015)
Under H_0, with high probability

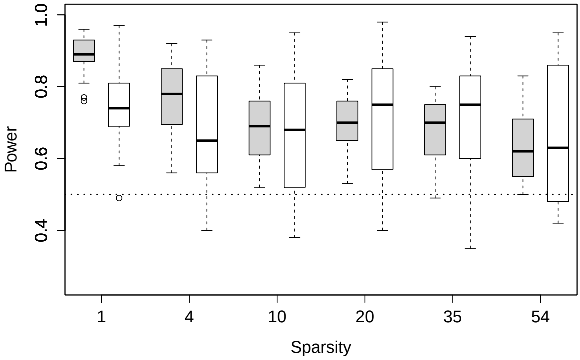
$$\hat{R}_\lambda \stackrel{\mathcal{D}}{=} R_\lambda$$

assuming

- beta-min assumption and (compatibility) restricted ℓ_1-eigenvalue condition for the design
  $\leadsto$ beta-min assumption is still there... but the result with "$=$" is rather strong

Low-dimensional with $p < n$

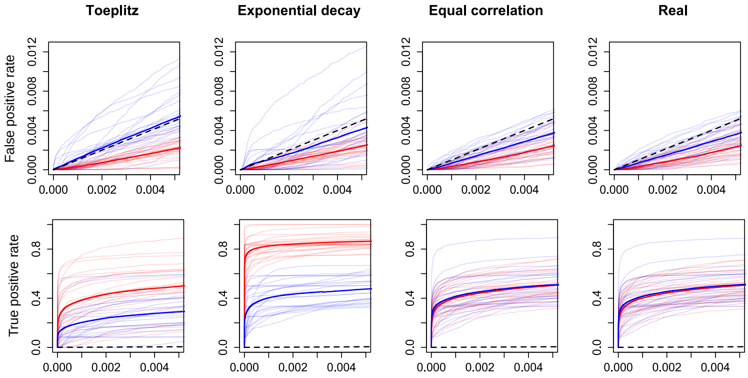
test whether 55 variables (corresponding to interactions and quadratic terms of 10 covariables) have no effect ($n = 442$; “diabetes dataset”)



- RP tests using Lasso (grey)
- Global test (Goeman et al., 2006) (white)
- F-test (dotted line)

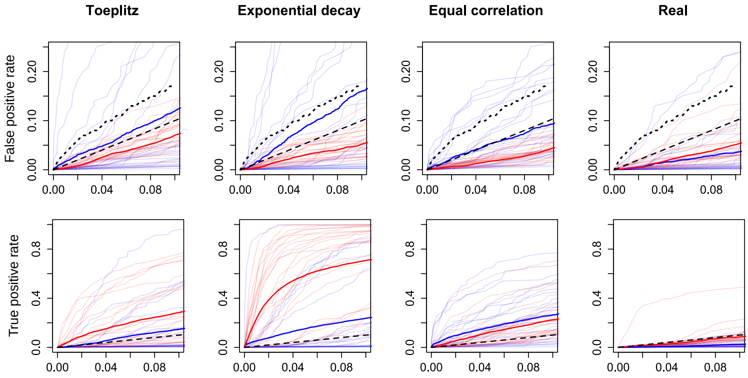
$\rightsquigarrow$ clearly more powerful than classical F-test!

Testing significance of individual variables



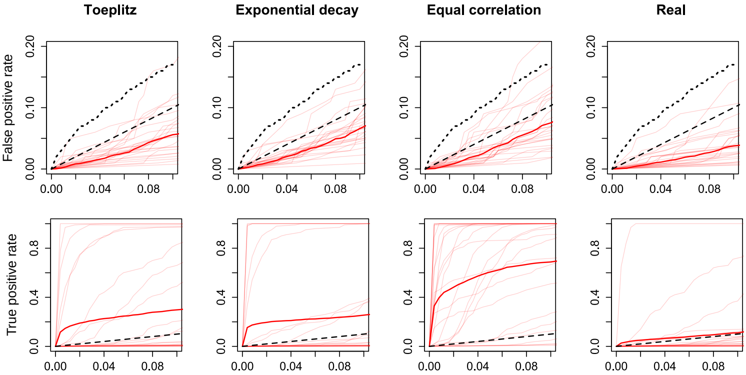
empirical distribution functions of *p*-values from RP tests and de-sparsified Lasso under the null (top row) and alternative (bottom row)

Testing significance of groups of variables



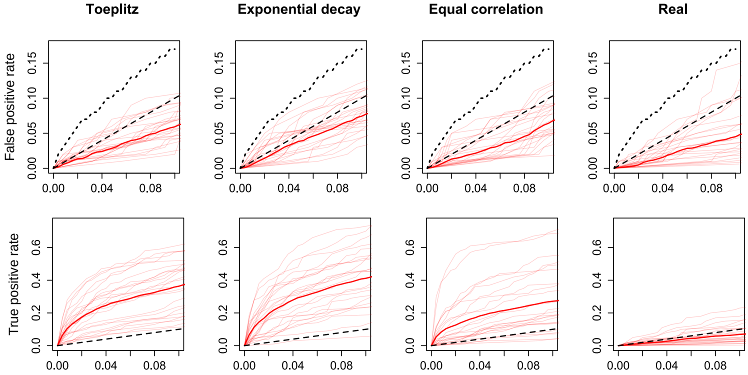
empirical distribution functions of *p*-values from RP tests and de-sparsified Lasso under the null (top row) and alternative (bottom row)

Testing for nonlinearity



RP method: Random Forests and OOB error as the proxy for prediction error

Testing for heteroscedasticity



RP method: regression of squared residuals using Lasso

$\rightsquigarrow$

RP testing "technology" can address some questions on "structural/model uncertainty" in high dimensions

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

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 (not needed for prediction)

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

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

model misspec. and uncertainty: RP test (Shah & PB, 2015)

inhomogeneous data (Meinshausen & PB, 2015; PB & Meinshausen, 2016) robustness, reliability and reproducibility of results...

in view of (yet) uncheckable assumptions

$\leadsto$

confirmatory high-dimensional inference
remains an interesting challenge

Thank you!

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.

- Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. Bernoulli 19, 1212-1242.
- 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.
- 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. (2013). Hierarchical testing in the high-dimensional setting with correlated variables. Journal of the American Statistical Association, published online (DOI: 10.1080/01621459.2015.1007209).
- Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E. and Bühlmann, P. (2015). Assessing statistical significance in joint analysis for genome-wide association studies. Bioinformatics, published online (DOI: 10.1093/bioinformatics/btw128).
- Mandozzi, J. and Bühlmann, P. (2015). A sequential rejection testing method for high-dimensional regression with correlated variables. To appear in International Journal of Biostatistics. Preprint arXiv:1502.03300
- Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. Electronic Journal of Statistics 9, 1449-1473.
- Shah, R.D. and Bühlmann, P. (2015). Goodness of fit tests for high-dimensional models. Preprint arXiv:1511.03334
- Meinshausen, N. and Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. Annals of Statistics 43, 1801-1830.
- Bühlmann, P. and Meinshausen, N. (2016). Magging: maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE 104, 126–135.