Some problems related to the Analysis of Large Dimensional Data

M. Bogdan

University of Wroclaw

Wroclaw, 17/06/2016

Outline

- Multiple testing

Outline

- Multiple testing
- Classical Model Selection Criteria for Multiple Regression

Outline

- Multiple testing
- Classical Model Selection Criteria for Multiple Regression
- Model Selection Selection Criteria Based on Convex Optimization

Outline

- Multiple testing
- Classical Model Selection Criteria for Multiple Regression
- Model Selection Selection Criteria Based on Convex Optimization
- SLOPE (Sorted L-One Penalized Estimation)

Classical example: Principle Components Analysis

Classical example: Principle Components Analysis

$X_{n \times p}$ - data matrix

Classical example: Principle Components Analysis

$X_{n\times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Analysis of Large Data

Classical example: Principle Components Analysis

$X_{n\times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Classical example: Principle Components Analysis

$X_{n \times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Purpose - understanding the biological/economical etc phenomena which generate the data, data compression (few basis vectors [principal components] may contain most of the information in the data), missing values imputation

Classical example: Principle Components Analysis

$X_{n\times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Purpose - understanding the biological/economical etc phenomena which generate the data, data compression (few basis vectors [principal components] may contain most of the information in the data), missing values imputation

General goal of large data analysis - separating the signal from noise, identifying the low dimensional structure spanning the noisy data

Analysis of Large Data

Classical example: Principle Components Analysis

$X_{n\times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Purpose - understanding the biological/economical etc phenomena which generate the data, data compression (few basis vectors [principal components] may contain most of the information in the data), missing values imputation

General goal of large data analysis - separating the signal from noise, identifying the low dimensional structure spanning the noisy data

Major problem - multiple comparisons, multiple testing (in PCA selection of nonzero singular values)

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - gene expressions of p genes for n_2 cancer patients

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - gene expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - gene expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - gene expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene j is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - gene expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene j is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

We test $H_{0j} : \mu_{1j} = \mu_{2j}$ with a t-test $t_j = \frac{\bar{X}_{.j} - \bar{Y}_{.j}}{S(\bar{X}_{.j} - \bar{Y}_{.j})}$, where $S(\bar{X}_{.j} - \bar{Y}_{.j})$ is the estimate of standard deviation of $\bar{X}_{.j} - \bar{Y}_{.j}$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - gene expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene j is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

We test $H_{0j} : \mu_{1j} = \mu_{2j}$ with a t-test $t_j = \frac{\bar{X}_{.j} - \bar{Y}_{.j}}{S(\bar{X}_{.j} - \bar{Y}_{.j})}$, where $S(\bar{X}_{.j} - \bar{Y}_{.j})$ is the estimate of standard deviation of $\bar{X}_{.j} - \bar{Y}_{.j}$

If n_1 and n_2 are large enough than $t_j \sim N(\mu_j, 1)$ with $\mu_j = \frac{\mu_{1j} - \mu_{2j}}{\sigma_{1j}/\sqrt{n_1} + \sigma_{2j}/\sqrt{n_2}}$ and $H_{0j} : \mu_j = 0$

Multiple testing (1)

$$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$$

Multiple testing (1)

$$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$$

$$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$$

Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0$ vs $\mu_i \neq 0$

Reject H_{0i} when $|X_i| > c$

Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0$ vs $\mu_i \neq 0$

Reject H_{0i} when $|X_i| > c$

Multiple comparison problem: if all μ_is are equal to zero than $max(|X_1|, \ldots, |X_p|) = \sqrt{2 \log p}(1 + o_p)$

Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

Reject H_{0i} when $|X_i| > c$

Multiple comparison problem: if all μ_is are equal to zero than $max(|X_1|, \ldots, |X_p|) = \sqrt{2 \log p}(1 + o_p)$

Thus to separate signal from noise we need $c = c(p) \to \infty$ as $p \to \infty$.

Multiple testing (2)

Significance level: $\alpha = P_{H_{0i}}(|X_i| > c)$

Multiple testing (2)

Significance level: $\alpha = P_{H_{0i}}(|X_i| > c)$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	m

Multiple testing (2)

Significance level: $\alpha = P_{H_{0i}}(|X_i| > c)$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	m

$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$

Multiple testing (2)

Significance level: $\alpha = P_{H_{0i}}(|X_i| > c)$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	m

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

Significance level: $\alpha = P_{H_{0i}}(|X_i| > c)$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	m

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

$$\alpha = 0.05, p_0 = 5000 \rightarrow E(V) = 250$$

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq = c(p) = \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)$

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq= c(p) = \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)$

$c(p) = \sqrt{2\log p}(1 + o_p)$

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq= c(p) = \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)$

$c(p) = \sqrt{2\log p}(1 + o_p)$

Benjamini-Hochberg procedure:

(1) $|X|_{(1)} \geq |X|_{(2)} \geq \ldots \geq |X|_{(p)}$

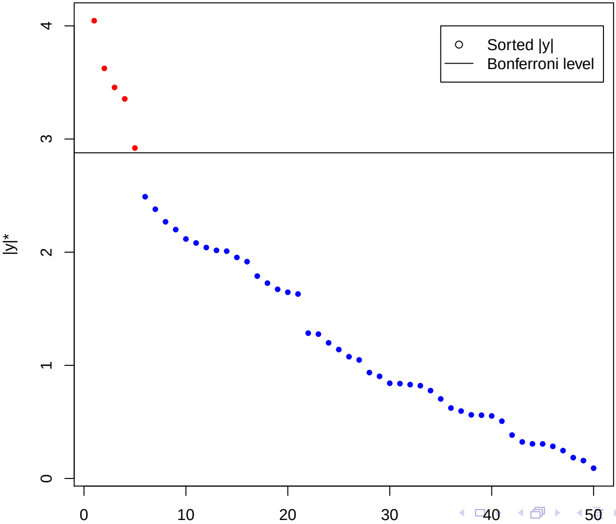
(2) Find the largest index i such that

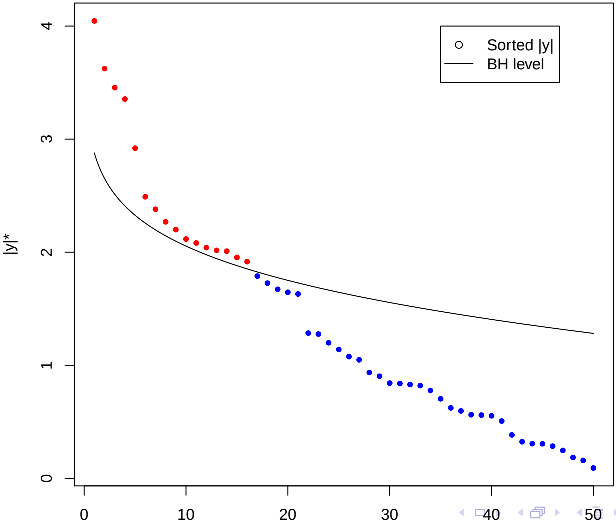
$$|X|_{(i)} \geq \Phi^{-1}(1 - \alpha_i), \quad \alpha_i = \alpha\frac{i}{2p}, \tag{1}$$

Call this index i_{SU}.

(3) Reject all $H_{(i)}$'s for which $i \leq i_{\mathrm{SU}}$

Bonferroni correction




Sorted |y|
BH level
|y|*
0
1
2
3
4
0
10
20
30
40
50

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR:

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}, \tag{2}$$

where p_0 is the number of true null hypotheses, $p_0 = |\{i : \mu_i = 0\}|$

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}, \tag{2}$$

where p_0 is the number of true null hypotheses, $p_0 = |\{i : \mu_i = 0\}|$

(Benjamini, Yekutieli, 2001) If test statistics are dependent then BH controls FDR at the level $\alpha\frac{p_0}{p}$ if $|X|_{(i)}$ is compared with $\Phi^{-1}\left(1 - \frac{i\alpha}{p\sum_{i=1}^{p}\frac{1}{i}}\right)$.

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}, \qquad (2)$$

where p_0 is the number of true null hypotheses, $p_0 = |\{i : \mu_i = 0\}|$

(Benjamini, Yekutieli, 2001) If test statistics are dependent then BH controls FDR at the level $\alpha \frac{p_0}{p}$ if $|X|_{(i)}$ is compared with $\Phi^{-1}\left(1 - \frac{i\alpha}{p \sum_{i=1}^{p} \frac{1}{i}}\right)$.

Detection thresholds:

for Bonferroni $\mu_i > (1+\epsilon)\sqrt{2 \log p)}$

for BH $\mu_i > (1+\epsilon)\sqrt{2(1-\beta) \log p}$, where the number of nonzero μ_i is proportional to p^{β} for some $\beta < 1$

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$
Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$
Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$
Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - classification problem

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$
Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$
Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - classification problem
Bayes risk, γ_0 - loss for type I error, γ_A - loss for type II error

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$
Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$
Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - classification problem
Bayes risk, γ_0 - loss for type I error, γ_A - loss for type II error
$\mu_j \sim (1-\epsilon)\delta_0 + \epsilon N(0, \tau^2)$

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity

$\epsilon \to 0$ as $p \to \infty$

Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$

Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - classification problem

Bayes risk, γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_j \sim (1-\epsilon)\delta_0 + \epsilon N(0, \tau^2)$

Bayes oracle $\to$ Bayes classifier, reject H_{0j} if $\frac{f_A(X_j)}{f_0(X_j)} > \frac{\gamma_0}{\gamma_A}\frac{1-\epsilon}{\epsilon}$, where $f_A(\cdot) = \phi(\cdot, 0, 1)$ and $f_0(\cdot) = \phi(\cdot, 0, 1+\tau^2)$

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$
Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$
Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - classification problem
Bayes risk, γ_0 - loss for type I error, γ_A - loss for type II error
$\mu_j \sim (1-\epsilon)\delta_0 + \epsilon N(0, \tau^2)$
Bayes oracle $\to$ Bayes classifier, reject H_{0j} if $\frac{f_A(X_j)}{f_0(X_j)} > \frac{\gamma_0}{\gamma_A}\frac{1-\epsilon}{\epsilon}$, where $f_A(\cdot) = \phi(\cdot, 0, 1)$ and $f_0(\cdot) = \phi(\cdot, 0, 1+\tau^2)$
The rule is Asymptotically Bayes Optimal under Sparsity (ABOS) if $\lim \frac{R}{R_{opt}} \to 1$, where R is the expected value of the loss (as $p \to \infty$)

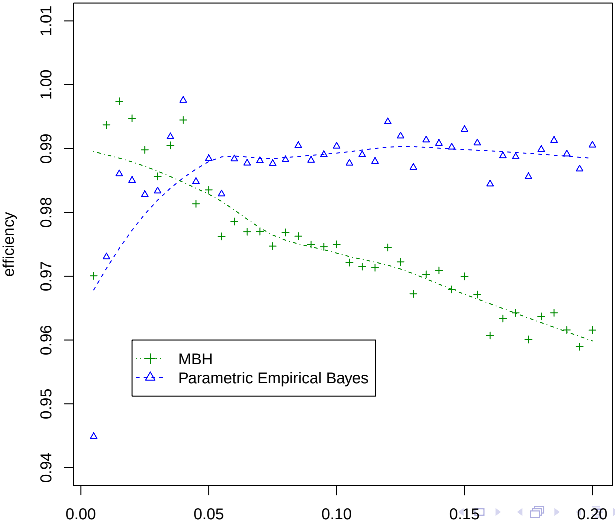
Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$
Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$
Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - classification problem
Bayes risk, γ_0 - loss for type I error, γ_A - loss for type II error
$\mu_j \sim (1-\epsilon)\delta_0 + \epsilon N(0,\tau^2)$
Bayes oracle $\to$ Bayes classifier, reject H_{0j} if $\frac{f_A(X_j)}{f_0(X_j)} > \frac{\gamma_0}{\gamma_A}\frac{1-\epsilon}{\epsilon}$, where $f_A(\cdot) = \phi(\cdot, 0, 1)$ and $f_0(\cdot) = \phi(\cdot, 0, 1+\tau^2)$
The rule is Asymptotically Bayes Optimal under Sparsity (ABOS) if $\lim \frac{R}{R_{opt}} \to 1$, where R is the expected value of the loss (as $p \to \infty$)
Bonferroni correction is ABOS if $\epsilon \approx \frac{1}{p}$

Asymptotic optimality of BH

$\epsilon = k/p$ - fraction of alternatives among all tests, sparsity
$\epsilon \to 0$ as $p \to \infty$
Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss : $||\hat{\mu} - \mu||$
Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - classification problem
Bayes risk, γ_0 - loss for type I error, γ_A - loss for type II error
$\mu_j \sim (1-\epsilon)\delta_0 + \epsilon N(0, \tau^2)$
Bayes oracle $\to$ Bayes classifier, reject H_{0j} if $\frac{f_A(X_j)}{f_0(X_j)} > \frac{\gamma_0}{\gamma_A}\frac{1-\epsilon}{\epsilon}$, where $f_A(\cdot) = \phi(\cdot, 0, 1)$ and $f_0(\cdot) = \phi(\cdot, 0, 1+\tau^2)$
The rule is Asymptotically Bayes Optimal under Sparsity (ABOS) if $\lim \frac{R}{R_{opt}} \to 1$, where R is the expected value of the loss (as $p \to \infty$)
Bonferroni correction is ABOS if $\epsilon \approx \frac{1}{p}$
BH is ABOS if $\epsilon \to 0$ and $k = p\epsilon \to C \in (0, \infty]$

Efficiency of BH with respect to misclassification probability



Finding predictors in Large Data Bases - Multiple Regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma^2 I)$$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma^2 I)$$

least squares estimator - $\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1})$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma^2 I)$$

least squares estimator - $\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1})$

Decide if $\beta_j = 0$ based on $z_j = \frac{\hat{\beta}_j}{(X'X)^{-1}_{jj}\sigma} \sim N\left(\frac{\beta_j}{(X'X)^{-1}_{jj}\sigma}, 1\right)$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma^2 I)$$

least squares estimator - $\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1})$

Decide if $\beta_j = 0$ based on $z_j = \frac{\hat{\beta}_j}{(X'X)^{-1}_{jj}\sigma} \sim N\left(\frac{\beta_j}{(X'X)^{-1}_{jj}\sigma}, 1\right)$

If p is large we need to deal with the multiple testing problem and with large variance of $\hat{\beta}$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma^2 I)$$

least squares estimator - $\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1})$

Decide if $\beta_j = 0$ based on $z_j = \frac{\hat{\beta}_j}{(X'X)^{-1}_{jj}\sigma} \sim N\left(\frac{\beta_j}{(X'X)^{-1}_{jj}\sigma}, 1\right)$

If p is large we need to deal with the multiple testing problem and with large variance of $\hat{\beta}$

Solution - Model Selection Criteria

Penalizing the model size

Classical model selection criteria

Miminimize

$$||Y - X\hat{\beta}||^2 + 2\sigma^2 \text{Penalty} \cdot k_M$$

Penalizing the model size

Classical model selection criteria

Miminimize

$$||Y - X\hat{\beta}||^2 + 2\sigma^2 \text{Penalty} \cdot k_M$$

Examples: AIC, BIC, RIC, Mallows C_P, etc.

AIC ... Penalty $= 1$, , BIC ... Penalty $= 1/2 \log n$, RIC ... Penalty$= \log p$

Penalizing the model size

Classical model selection criteria

Miminimize

$$||Y - X\hat{\beta}||^2 + 2\sigma^2 \text{Penalty} \cdot k_M$$

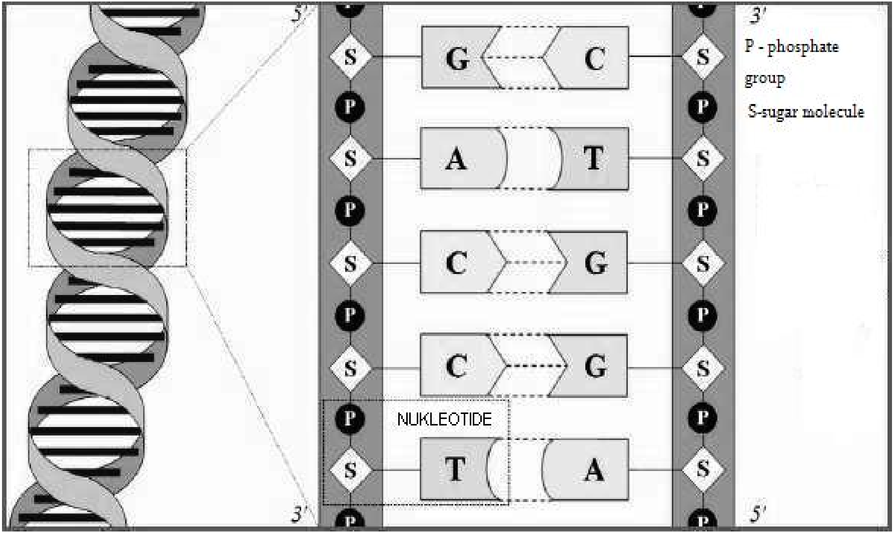
Examples: AIC, BIC, RIC, Mallows C_P, etc.

AIC ... Penalty $= 1$, , BIC ... Penalty $= 1/2 \log n$, RIC ... Penalty$= \log p$

Problem - combinatorial search, heuristic search procedures.

Solution - Convex optimization framework

DNA structure



Genetic variability

- About 99,9% of genetic information is the same for all people.
- A ***polymorphism*** is a difference in DNA structure, which is present in at least 1% of population
- A ***Single Nucleotide Polymorphism(SNP)*** is a polymorphism with the difference in the single base:
  - A typical SNP: a position in DNA in which
    - 85% of population has Cytosine(C)
    - 15% has a Thymine(T).
- There are usually two forms of a SNP at a given locus
- three genotypes : AA, Aa, aa.

Main purpose

MAIN PURPOSE: finding the mutations in DNA sequence, that influence the trait of interest.

Main purpose

MAIN PURPOSE: finding the mutations in DNA sequence, that influence the trait of interest.

Y - quantitative trait

Main purpose

MAIN PURPOSE: finding the mutations in DNA sequence, that influence the trait of interest.

Y - quantitative trait

Examples: blood pressure, cholesterol level, gene expression level

Data structure

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

Data structure

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

$X_{n \times p}$ - matrix of SNP genotypes

Data structure

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

$X_{n \times p}$ - matrix of SNP genotypes

Usually $n \approx k \times 100$ or $k \times 1000$, $p \approx k \times 10,000$ or $p \approx k \times 100,000$

Statistical model

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, I)$$

Statistical model

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, I)$$

Goals:

1. Identification of nonzero elements in vector of regression coefficients β.
2. Building a good predictive model - minimizing $|\hat{\beta} - \beta|$

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, I)$$

Goals:

1. Identification of nonzero elements in vector of regression coefficients β.
2. Building a good predictive model - minimizing $|\hat{\beta} - \beta|$

Generalizations:

a) adding nonlinear terms and interactions
b) Generalized Linear Models

$$E(Y) = G^{-1}(X\beta), \quad \text{G - link function}$$

E.g. Binary Y - logistic regression (e.g. identification of factors influencing the credit risk or the risk of developing some disease)

Statistical model

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, I)$$

Goals:

1. Identification of nonzero elements in vector of regression coefficients β.
2. Building a good predictive model - minimizing $|\hat{\beta} - \beta|$

Generalizations:

a) adding nonlinear terms and interactions

b) Generalized Linear Models

$$E(Y) = G^{-1}(X\beta), \quad \text{G - link function}$$

E.g. Binary Y - logistic regression (e.g. identification of factors influencing the credit risk or the risk of developing some disease)

General class of problems - identifying important factors when looking through large data bases

No noise case - linear program

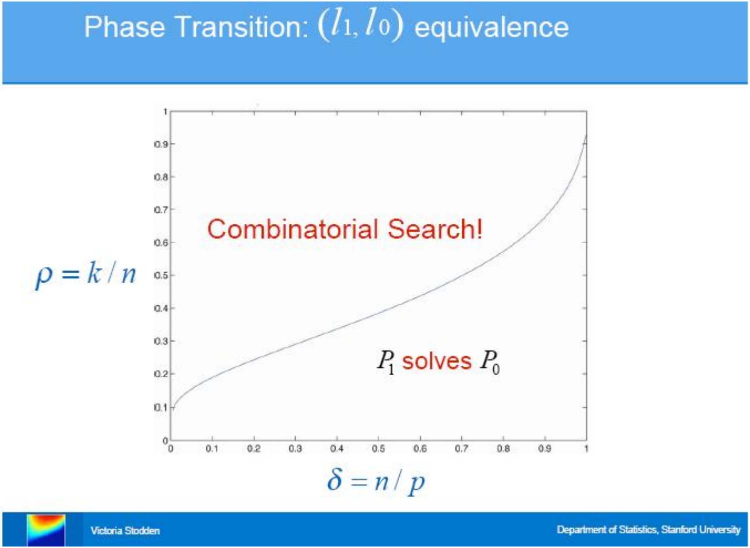
$$Y = X\beta$$

No noise case - linear program

$$Y = X\beta$$

If $p > n$ minimize $||\beta||_1 = \sum_{i=1}^{n} |\beta_i|$ subject to $Y = X\beta$.

Transition curve, Donoho and Tanner (2005), X is Gaussian



Columns in general position

Points $X_1, \ldots, X_p \in R^n$ are said to be in general position provided that the affine span of any $k+1$ points $s_1 X_{i_1}, \ldots, s_{k+1} X_{i_{k+1}}$, for any any signs $s_1, \ldots, s_{k+1} \in \{-1, 1\}$, does not contain any element of the set $\{\pm X_i, i \neq i_1, \ldots, i_{k+1}\}$.

Transition curve (2)

Cross-polytope:

$$C^p := \left\{ \beta \in R^p : \sum_{i=1}^{p} |\beta_i| \leq 1 \right\}$$

Transition curve (2)

Cross-polytope:

$$C^p := \left\{ \beta \in R^p : \sum_{i=1}^{p} |\beta_i| \leq 1 \right\}$$

Theorem

Let X be a fixed matrix with p columns in general position in R^n. Consider vectors y_0 with a sparse solution $y_0 = X\beta_0$, where β_0 has k nonzeros. The fraction of systems (y_0, X) where the convex program has that underlying β_0 as its unique solution is $f_k(XC^p)/f_k(C^p)$, where $f_k(\cdot)$ is the number of k dimensional faces of the polytope.

Transition curve (2)

Cross-polytope:

$$C^p := \left\{ \beta \in R^p : \sum_{i=1}^{p} |\beta_i| \leq 1 \right\}$$

Theorem

Let X be a fixed matrix with p columns in general position in R^n. Consider vectors y_0 with a sparse solution $y_0 = X\beta_0$, where β_0 has k nonzeros. The fraction of systems (y_0, X) where the convex program has that underlying β_0 as its unique solution is $f_k(XC^p)/f_k(C^p)$, where $f_k(\cdot)$ is the number of k dimensional faces of the polytope.

Let's denote $\rho = k/p$ and $\delta = n/p$. For the Gaussian matrix X $\lim_{p\to\infty} f_k(XC^p)/f_k(C^p) = 1$ if $\rho < \rho(\delta)$ and 0 if $\rho > \rho(\delta)$.

Noisy case - statistical problem

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Noisy case - statistical problem

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \epsilon$

Or alternatively: $\min_{b\in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda\sigma||b||_1$

Noisy case - statistical problem

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \epsilon$

Or alternatively: $\min_{b\in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda\sigma||b||_1$

In statistics this procedure is called LASSO (Tibshirani, 1996)

Some theoretical results: Candes, Plan (2008)

Assumption, notation:

a) for every $i \in \{1, \ldots, p\}$ $||X_i||_2 = 1$

b) We denote $\mu(X) = \sup_{1 \le i < j \le p} | < X_i, X_j > |$

Some theoretical results: Candes, Plan (2008)

Assumption, notation:

a) for every $i \in \{1, \ldots, p\}$ $||X_i||_2 = 1$

b) We denote $\mu(X) = \sup_{1 \leq i < j \leq p} | < X_i, X_j > |$

A matrix X is said to obey the coherence property if

$$\mu(X) \leq A_0 (\log p)^{-1} \quad .$$

Some theoretical results: Candes, Plan (2008)

Assumption, notation:

a) for every $i \in \{1, \dots, p\}$ $||X_i||_2 = 1$
b) We denote $\mu(X) = \sup_{1 \leq i < j \leq p} | < X_i, X_j > |$

A matrix X is said to obey the coherence property if

$$\mu(X) \leq A_0 (\log p)^{-1} \quad .$$

Eg. if $x_{ij} \sim N(0, 1/n)$ then $\mu(X) \sim \sqrt{2 \log p/n}$

Asymptotically optimal prediction

Theorem

Suppose that X obeys the coherence property and assume that $||\beta||_0 \leq S \leq c_0 p/[||X||^2 logp]$. Then the lasso estimate computed with $\lambda = 2\sqrt{2\log p}$ obeys

$$||X\beta - X\hat{\beta}||_2^2 \leq C_0(2\log p)S\sigma^2 \quad ,$$

with probability at least $1 - 6p^{-2\log 2} - p^{-1}(2\pi \log p)^{-1/2}$. The constant C_0 may be taken as $8(1+\sqrt{2})^2$.

For Gaussian design $||X||^2 \sim \sqrt{p/n}$ so $S \leq c_0 n/\log p$.

Exact model recovery

Theorem

Let I be the support of β and suppose that

$$\min_{i \in I} |\beta_i| > 8\sigma\sqrt{2\log p}$$

. Then under the above assumption the lasso estimate with $\lambda = 2\sqrt{2\log p}$ *obeys*

$$supp(\hat{\beta}) = supp(\beta) \ \ and$$

$$sgn(\hat{\beta}_i) = sgn(\beta_i) \ \ for\ all \ \ i \in I$$

with probability at least
$1 - 2p^{-1}((2\pi\log p)^{-1/2} + |I|p^{-1}) - O(p^{-2\log 2}$.

Our goal

Goal - Construction of the procedure with the finite sample statistical guarantees like e.g. control of the false discovery rate (FDR)

LASSO, (Tibshirani, 1996)

$$\min_{b \in \mathbb{R}^m} \left(\frac{1}{2} \|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1} \right). \tag{3}$$

LASSO, (Tibshirani, 1996)

$$\min_{b \in \mathbb{R}^m} \left(\frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1}\right). \tag{3}$$

LASSO solution

$$\hat{\beta} = \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - y)) = \eta_\lambda(\hat{\beta} - X'X(\hat{\beta} - \beta) + X'z) \quad ,$$

where $\eta_\lambda(t) = \mathrm{sgn}(t)(|t| - \lambda)_+$, applied componentwise

LASSO, (Tibshirani, 1996)

$$\min_{b \in \mathbb{R}^m} \left(\tfrac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1}\right). \quad (3)$$

LASSO solution

$$\hat{\beta} = \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - y)) = \eta_\lambda(\hat{\beta} - X'X(\hat{\beta} - \beta) + X'z) \ ,$$

where $\eta_\lambda(t) = \mathrm{sgn}(t)(|t| - \lambda)_+$, applied componentwise

When $X'X = I$ and $z \sim N(0, \sigma)$

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i), \quad (4)$$

where $Z_i \sim N(0, \sigma)$.

If $\beta = 0$ then ($\hat{\beta} > 0$ if $|Z_i| > \lambda$)

$$\min_{b \in \mathbb{R}^m} \left(\tfrac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1}\right). \tag{3}$$

LASSO solution

$$\hat{\beta} = \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - y)) = \eta_\lambda(\hat{\beta} - X'X(\hat{\beta} - \beta) + X'z) \quad ,$$

where $\eta_\lambda(t) = \mathrm{sgn}(t)(|t| - \lambda)_+$, applied componentwise
When $X'X = I$ and $z \sim N(0, \sigma)$

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i), \tag{4}$$

where $Z_i \sim N(0, \sigma)$.
If $\beta = 0$ then ($\hat{\beta} > 0$ if $|Z_i| > \lambda$)
$\lambda = \sigma\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \approx \sigma\sqrt{2\log p}$ - Bonferroni correction

LASSO, (Tibshirani, 1996)

$$\min_{b \in \mathbb{R}^m} \left(\tfrac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1}\right). \tag{3}$$

LASSO solution

$$\hat{\beta} = \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - y)) = \eta_\lambda(\hat{\beta} - X'X(\hat{\beta} - \beta) + X'z) \quad ,$$

where $\eta_\lambda(t) = \mathrm{sgn}(t)(|t| - \lambda)_+$, applied componentwise

When $X'X = I$ and $z \sim N(0, \sigma)$

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i), \tag{4}$$

where $Z_i \sim N(0, \sigma)$.

If $\beta = 0$ then ($\hat{\beta} > 0$ if $|Z_i| > \lambda$)

$\lambda = \sigma\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \approx \sigma\sqrt{2\log p}$ - Bonferroni correction

Nonadaptive - relatively low power

Sorted L-One Penalized Estimation: SLOPE

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0, \tag{5}$$

Sorted L-One Penalized Estimation: SLOPE

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0, \tag{5}$$

Sorted L_1 norm:

$$J_\lambda(b) = \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)} + \ldots + \lambda_p|b|_{(p)} \quad . \tag{6}$$

$J_\lambda(b)$ is convex because by the Hardy-Littlewood-Pólya inequality

$$J_\lambda(b) = \max_\pi \sum_{i=1}^{p} \lambda_{\pi(i)}|b_i| \quad ,$$

where the maximum is over all permutations of the elements of b.

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0, \tag{5}$$

Sorted L_1 norm:

$$J_\lambda(b) = \lambda_1 |b|_{(1)} + \lambda_2 |b|_{(2)} + \ldots + \lambda_p |b|_{(p)} \quad . \tag{6}$$

$J_\lambda(b)$ is convex because by the Hardy-Littlewood-Pólya inequality

$$J_\lambda(b) = \max_\pi \sum_{i=1}^{p} \lambda_{\pi(i)} |b_i| \quad ,$$

where the maximum is over all permutations of the elements of b.

SLOPE:

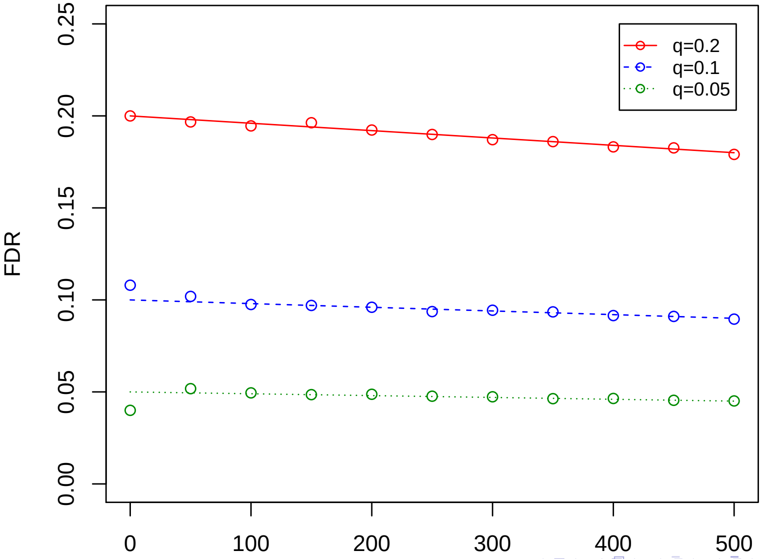
$$\min_{b \in \mathbb{R}^m} \ \frac{1}{2} \|y - Xb\|_{\ell_2}^2 + \sigma J_\lambda(b). \tag{7}$$

Theorem

Assume an orthogonal design with iid $\mathcal{N}(0, \sigma^2)$ errors, and set $\lambda_{BH}(i) = \Phi^{-1}(1 - iq/2p)$. Then the FDR of SLOPE obeys

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] \leq q\frac{m_0}{m}. \tag{8}$$

FDR Orthogonal design



Multiple testing - ANOVA (1)

1000 tests in 5 different laboratories

$$y_{i,j} = \mu_i + \tau_j + z_{i,j}, \quad 1 \leq i \leq 1000, \ 1 \leq j \leq 5,$$

,

$$\tau_j \sim N(0, \sigma_\tau^2), \ \ z_{i,j} \sim N(0, \sigma_z^2)$$

Multiple testing - ANOVA (1)

1000 tests in 5 different laboratories

$$y_{i,j} = \mu_i + \tau_j + z_{i,j}, \quad 1 \le i \le 1000,\ 1 \le j \le 5,$$

,

$$\tau_j \sim N(0, \sigma_\tau^2),\ \ z_{i,j} \sim N(0, \sigma_z^2)$$

$$\bar{y}_i = \mu_i + \bar{\tau} + \bar{z}_i, \quad 1 \le i \le 1000.$$

1000 tests in 5 different laboratories

$$y_{i,j} = \mu_i + \tau_j + z_{i,j}, \quad 1 \leq i \leq 1000, \; 1 \leq j \leq 5,$$

,

$$\tau_j \sim N(0, \sigma_\tau^2), \;\; z_{i,j} \sim N(0, \sigma_z^2)$$

$$\bar{y}_i = \mu_i + \bar{\tau} + \bar{z}_i, \quad 1 \leq i \leq 1000.$$

When $\sigma_\tau^2 = \sigma_z^2 = 2.5$ then

$$\bar{Y} \sim N(\mu, \Sigma)$$

$\Sigma_{ii} = \sigma = 1$ and $\Sigma_{ij} = 0.5$ for $i \neq j$.

1000 tests in 5 different laboratories

$$y_{i,j} = \mu_i + \tau_j + z_{i,j}, \quad 1 \leq i \leq 1000, \; 1 \leq j \leq 5,$$

,

$$\tau_j \sim N(0, \sigma_\tau^2), \;\; z_{i,j} \sim N(0, \sigma_z^2)$$

$$\bar{y}_i = \mu_i + \bar{\tau} + \bar{z}_i, \quad 1 \leq i \leq 1000.$$

When $\sigma_\tau^2 = \sigma_z^2 = 2.5$ then

$$\bar{Y} \sim N(\mu, \Sigma)$$

$\Sigma_{ii} = \sigma = 1$ and $\Sigma_{ij} = 0.5$ for $i \neq j$.

Goal - testing $H_{0i} : \mu_i = 0, \; i = 1, \ldots, 1000$ vs $H_{Ai} : \mu_i \neq 0$.

Multiple testing (2)

Marginal tests with BH: Compare $|\bar{y}|_{(i)}$ with $\sigma\Phi^{-1}(1 - \frac{\alpha i}{1000})$

Multiple testing (2)

Marginal tests with BH: Compare $|\bar{y}|_{(i)}$ with $\sigma\Phi^{-1}(1-\frac{\alpha i}{1000})$

Alternatively:

$$Y^\star = \Sigma^{-1/2}Y = \Sigma^{-1/2}\mu + \epsilon, \tag{9}$$

where $\epsilon \sim N(0, I_{p\times p})$

$U = \Sigma^{-1/2}$ has a dominating diagonal

$U(i,i) = 1.4128$ and $U(i,j) = -0.0014$.

Multiple testing (2)

Marginal tests with BH: Compare $|\bar{y}|_{(i)}$ with $\sigma\Phi^{-1}(1-\frac{\alpha i}{1000})$

Alternatively:

$$Y^\star = \Sigma^{-1/2}Y = \Sigma^{-1/2}\mu + \epsilon, \tag{9}$$

where $\epsilon \sim N(0, I_{p\times p})$

$U = \Sigma^{-1/2}$ has a dominating diagonal

$U(i,i) = 1.4128$ and $U(i,j) = -0.0014$.

Use SLOPE with λ_{BH} to identify nonzero elements of μ

Multiple testing (2)

Marginal tests with BH: Compare $|\bar{y}|_{(i)}$ with $\sigma\Phi^{-1}(1 - \frac{\alpha i}{1000})$

Alternatively:

$$Y^\star = \Sigma^{-1/2} Y = \Sigma^{-1/2}\mu + \epsilon, \qquad (9)$$

where $\epsilon \sim N(0, I_{p\times p})$

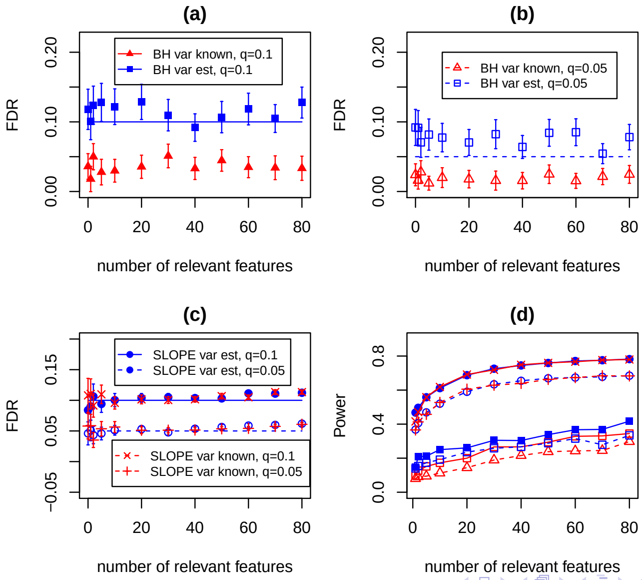
$U = \Sigma^{-1/2}$ has a dominating diagonal

$U(i,i) = 1.4128$ and $U(i,j) = -0.0014$.

Use SLOPE with λ_{BH} to identify nonzero elements of μ

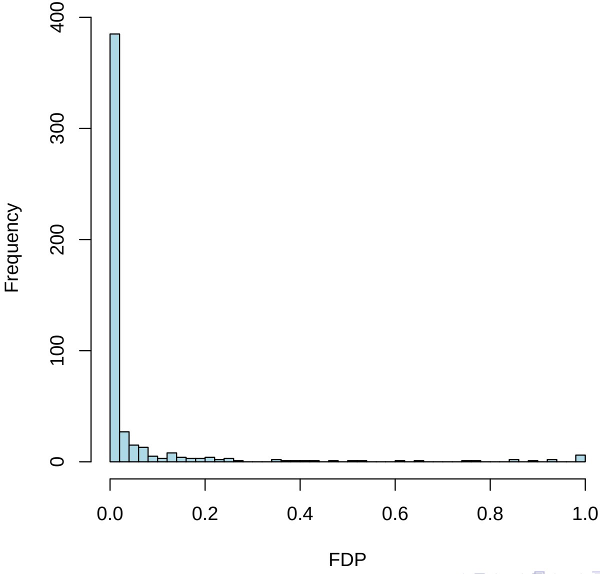
Unknown variance components: σ^2_τ and σ^2_z are estimated using classical unweighted means method

FDR

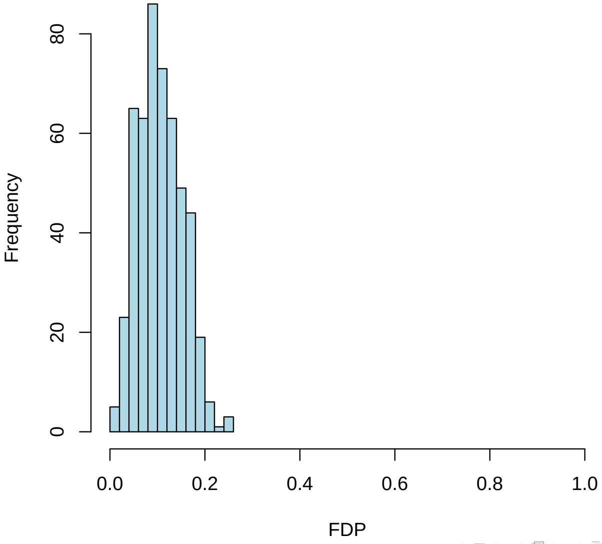


(a)
BH var known, q=0.1
BH var est, q=0.1
FDR
0.00
0.10
0.20
0
20
40
60
80
number of relevant features
(b)
BH var known, q=0.05
BH var est, q=0.05
FDR
0.00
0.10
0.20
0
20
40
60
80
number of relevant features
(c)
SLOPE var est, q=0.1
SLOPE var est, q=0.05
SLOPE var known, q=0.1
SLOPE var known, q=0.05
FDR
−0.05
0.05
0.15
0
20
40
60
80
number of relevant features
(d)
Power
0.0
0.4
0.8
0
20
40
60
80
number of relevant features

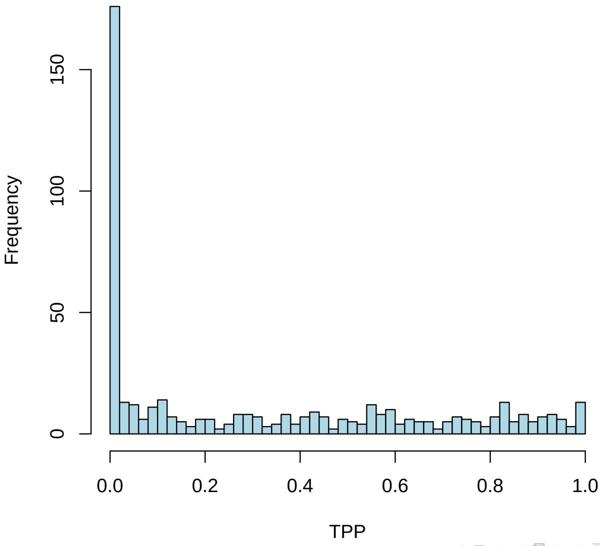
Distribution of FDP for marginal tests, $k = 50$



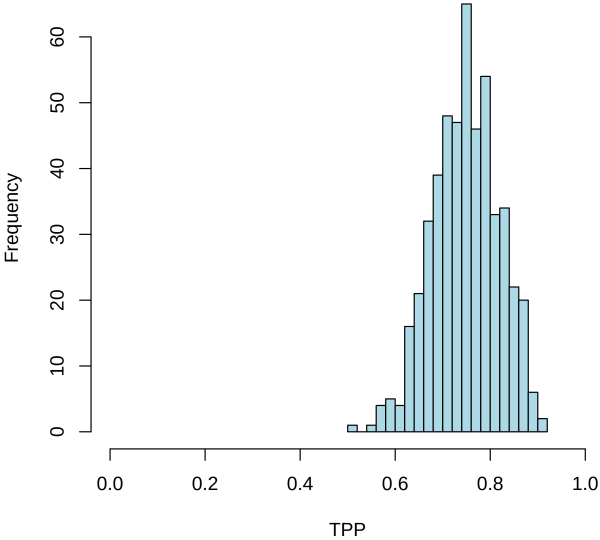
Distribution of FDP for SLOPE, $k = 50$



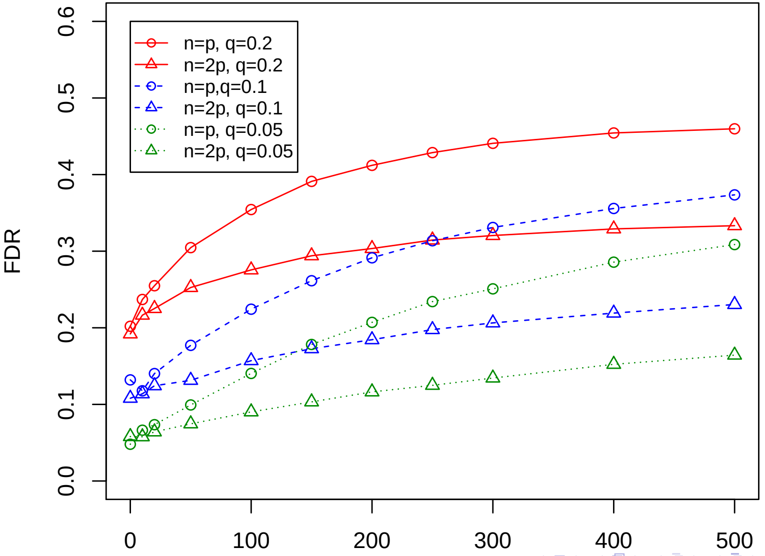
Distribution of TPP for marginal tests, $k = 50$



Distribution of TPP for SLOPE, $k = 50$



FDR Nonorthogonal design - SLOPE



Explanation

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i + v_i),$$

Explanation

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i + v_i),$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j) \rangle,$$

Explanation

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i + v_i),$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j)\rangle,$$

S - support of the true model, $|S| = k$

Assume that all signals are strong enough so they are detected by SLOPE. Then the respective regression coefficients

Explanation

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i + v_i),$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j)\rangle,$$

S - support of the true model, $|S| = k$

Assume that all signals are strong enough so they are detected by SLOPE. Then the respective regression coefficients

$$\hat{\beta}_S \approx (X_S' X_S)^{-1}(X_S' y - \lambda_S) = \hat{\beta}_{OLS} - (X_S' X_S)^{-1}\lambda_S,$$

where $\lambda_S = (\lambda(1), \ldots, \lambda(k))'$.

Explanation

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i + v_i),$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j) \rangle,$$

S - support of the true model, $|S| = k$

Assume that all signals are strong enough so they are detected by SLOPE. Then the respective regression coefficients

$$\hat{\beta}_S \approx (X_S' X_S)^{-1}(X_S' y - \lambda_S) = \hat{\beta}_{OLS} - (X_S' X_S)^{-1}\lambda_S,$$

where $\lambda_S = (\lambda(1), \ldots, \lambda(k))'$.

$$v_i \approx E X_i' X_S (X_S' X_S)^{-1} \lambda_S \quad .$$

Explanation

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i + v_i),$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j)\rangle,$$

S - support of the true model, $|S| = k$

Assume that all signals are strong enough so they are detected by SLOPE. Then the respective regression coefficients

$$\hat{\beta}_S \approx (X_S' X_S)^{-1}(X_S' y - \lambda_S) = \hat{\beta}_{OLS} - (X_S' X_S)^{-1}\lambda_S,$$

where $\lambda_S = (\lambda(1), \ldots, \lambda(k))'$.

$$v_i \approx E X_i' X_S (X_S' X_S)^{-1} \lambda_S \quad .$$

For gaussian matrices $x_{ij} \sim N(0, 1/n)$

$$\mathbb{E}(X_i' X_S (X_S' X_S)^{-1} \lambda_S)^2 = w(|\mathcal{S}|) \cdot \|\lambda_S\|_{\ell_2}^2, \quad w(k) = \frac{1}{n - k - 1}$$

Corrected version

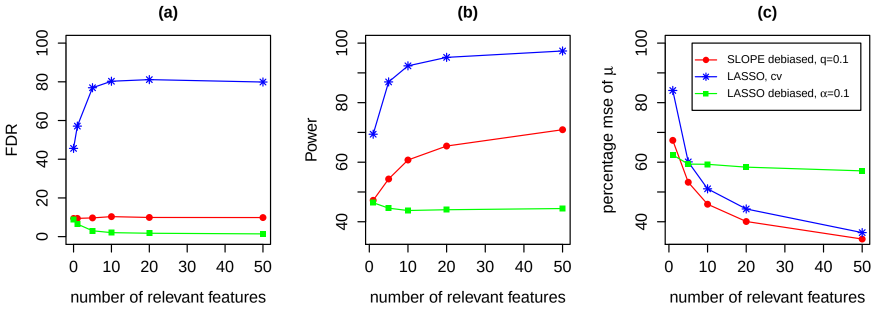
$$\lambda_G(i) = \lambda_{BH}(i)\sqrt{1 + w(i-1)\sum_{j<i}\lambda_G(j)^2}.$$

Corrected version

$$\lambda_G(i) = \lambda_{BH}(i)\sqrt{1 + w(i-1)\sum_{j<i}\lambda_G(j)^2}.$$

For other designs we estimate $E(X_i' X_S (X_S' X_S)^{-1}\lambda_S)^2$ by randomly drawing columns of the design matrix

FDR, $p = n = 5000$, Gaussian design



Unkown σ - scaled SLOPE

Algorithm 1 Iterative SLOPE fitting when σ is unknown

```
input: y, X and initial sequence λ^S
    (computed for σ = 1)
initialize: S_+ = ∅
repeat
    S = S_+
    compute the RSS obtained by regressing y onto variables in S
    set σ̂^2 = RSS/(n − |S| − 1)
    compute the solution β̂ to SLOPE with parameter sequence
        σ̂ · λ^S
    set S_+ = supp(β̂)
until S_+ = S
```

Simulation example

For $n = p = 5000$ we simulate 5000 genotypes of p independent Single Nucleotide Polymorphisms (SNPs)

Simulation example

For $n = p = 5000$ we simulate 5000 genotypes of p independent Single Nucleotide Polymorphisms (SNPs)

Scenario 1: $Y = X\beta + z$ - ideal linear model, only additive effects

Simulation example

For $n = p = 5000$ we simulate 5000 genotypes of p independent Single Nucleotide Polymorphisms (SNPs)

Scenario 1: $Y = X\beta + z$ - ideal linear model, only additive effects

Nonlinearity, dominance effects:

$$\tilde{z}_{ij} = \begin{cases} -1 & \text{for} \quad aa, AA \\ 1 & \text{for} \quad aA \end{cases}, \tag{10}$$

$$y = [X, Z][\beta_X', \beta_Z']' + \epsilon$$

The search is performed only over X matrix.

Simulation example

For $n = p = 5000$ we simulate 5000 genotypes of p independent Single Nucleotide Polymorphisms (SNPs)

Scenario 1: $Y = X\beta + z$ - ideal linear model, only additive effects

Nonlinearity, dominance effects:

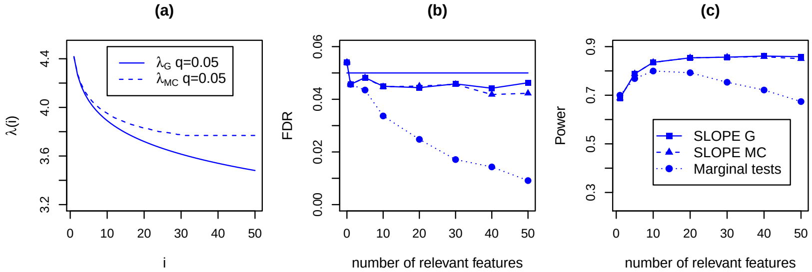
$$\tilde{z}_{ij} = \begin{cases} -1 & \text{for} \quad aa, AA \\ 1 & \text{for} \quad aA \end{cases}, \qquad (10)$$

$$y = [X, Z][\beta_X', \beta_Z']' + \epsilon$$

The search is performed only over X matrix.

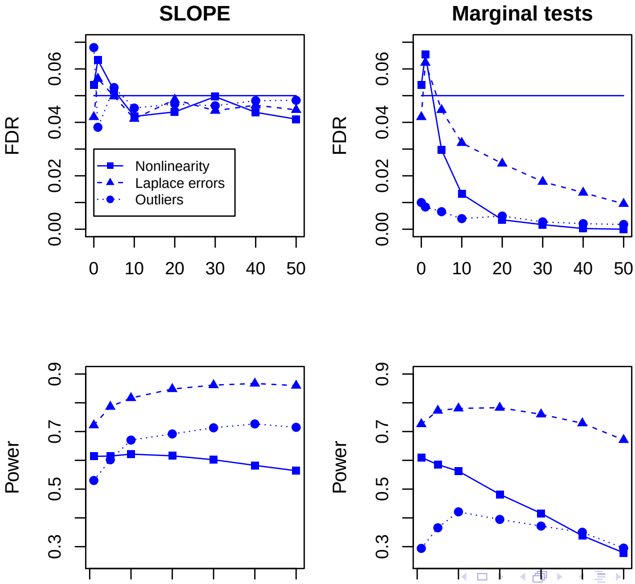
Errors from Laplace distribution and with some proportion of outliers

Ideal model



(a)
(b)
(c)
λ(i)
i
λ_G q=0.05
λ_{MC} q=0.05
FDR
Power
number of relevant features
number of relevant features
SLOPE G
SLOPE MC
Marginal tests

Violations of model assumptions



Real data analysis

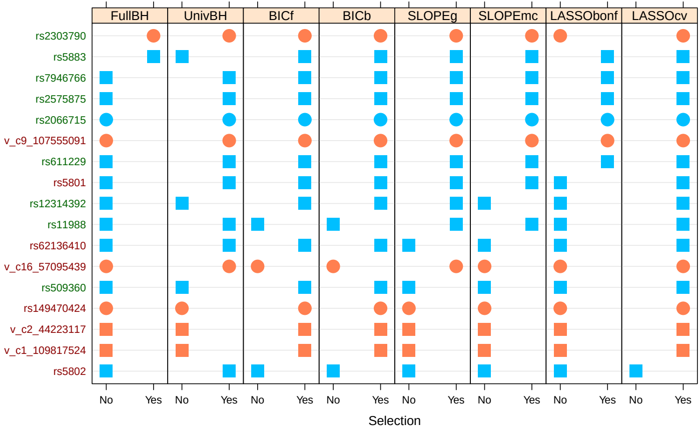
Y- fasting blood HDL levels

X - genotypes of 777 SNPs in interesting genome regions

$n = 5375$ individuals

maximal pairwise correlation between SNPs $= 0.3$

Results



FullBH
UnivBH
BICf
BICb
SLOPEg
SLOPEmc
LASSObonf
LASSOcv
rs2303790
rs5883
rs7946766
rs2575875
rs2066715
v_c9_107555091
rs611229
rs5801
rs12314392
rs11988
rs62136410
v_c16_57095439
rs509360
rs149470424
v_c2_44223117
v_c1_109817524
rs5802
No
Yes
Selection

Compressed sensing examples

$X_{n\times p}$ - selection of n rows from the one-dimensional discrete cosine transformation matrix, $n = p/2$, $p = 262{,}144$

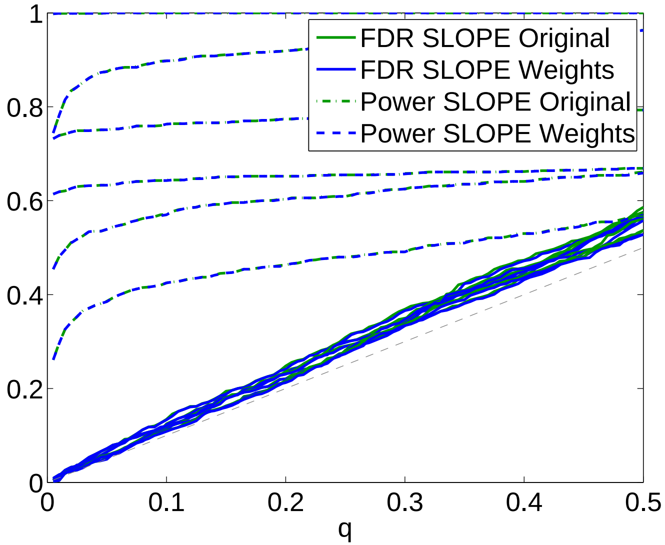
Compressed sensing examples

$X_{n\times p}$ - selection of n rows from the one-dimensional discrete cosine transformation matrix, $n = p/2$, $p = 262{,}144$

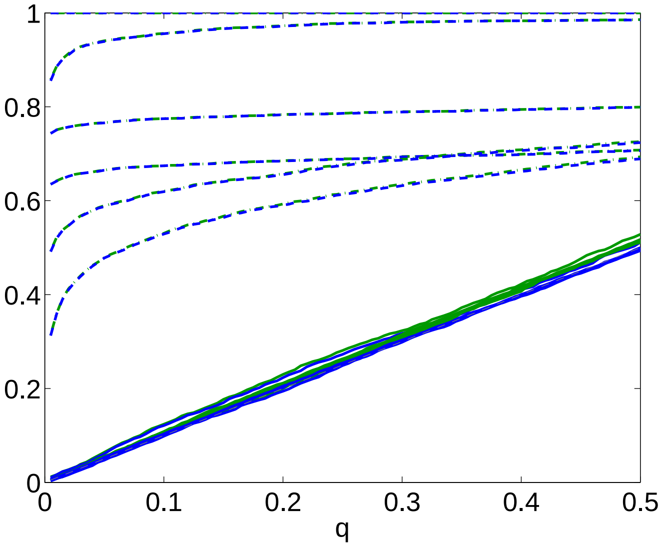
Signals:

1. Random Gaussian entries: $\beta_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 2\sqrt{2\log p}$.
2. Random Gaussian entries: $\beta_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 3\sqrt{2\log p}$.
3. Constant values: $\beta_i = 1.2\sqrt{2\log p}$.
4. Linearly decreasing from $1.2\sqrt{2\log p}$ to $0.6\sqrt{2\log p}$.
5. Linearly decreasing from $1.5\sqrt{2\log p}$ to $0.5\sqrt{2\log p}$.
6. Linearly decreasing from $4.5\sqrt{2\log p}$ to $1.5\sqrt{2\log p}$.
7. Exponentially decaying entries: $v_i = 1.2\sqrt{2\log p}\,(i/k)^{-1.2}$.

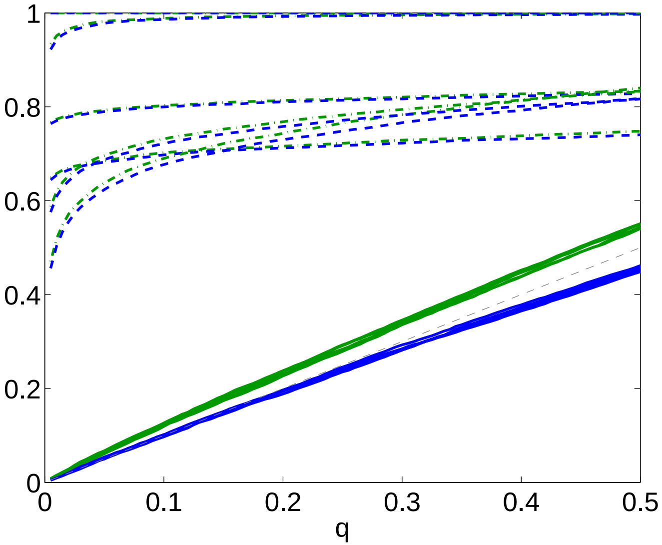
FDR of SLOPE, $k = 10$



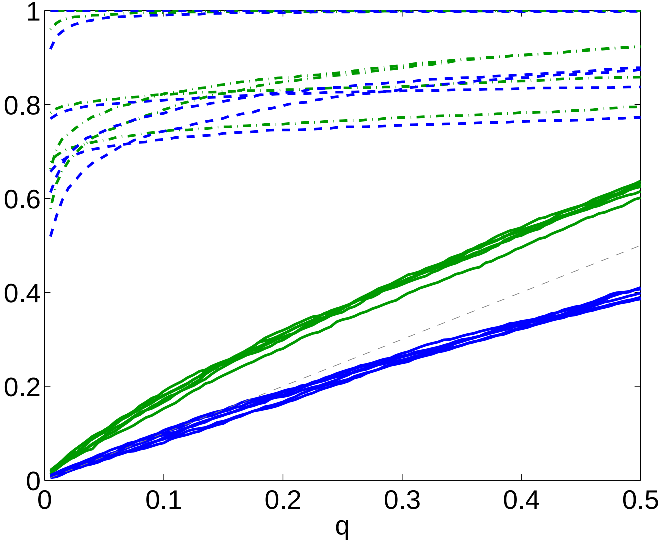
FDR of SLOPE, $k = 50$



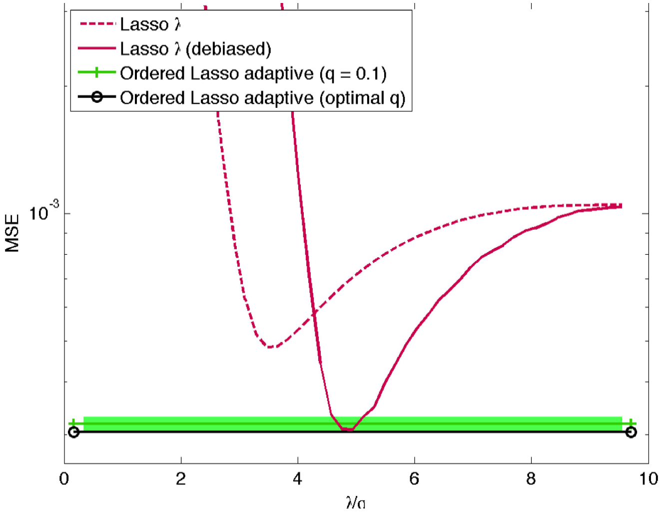
FDR of SLOPE, $k = 500$



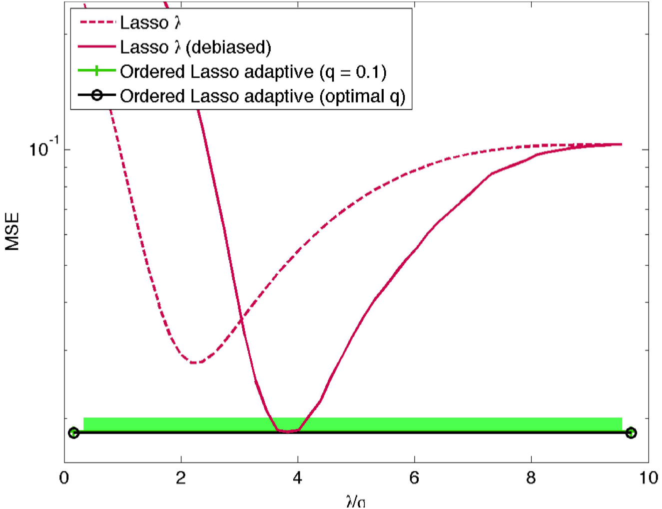
FDR of SLOPE, $k = 2621$



MSE, linearly decaying (5.3), $k = 10$



MSE, linearly decaying (5.7), $k = 1000$



MSE, SLOPE, linearly decaying

MSE, SLOPE, exponentially decaying

Many open problems

1. Proof of FDR control for random designs - possibly better choice of the regularizing sequence.
2. Asymptotic minimaxity of SLOPE
3. Universality of gaussian weights
4. Identification of the class of the covariance matrices for which SLOPE might be useful in the context of multiple testing.
5. Application for full GWAS studies.
6. Group SLOPE and Ordered Dantzig selector.
7. Other goals, e.g. see OSCAR (Biometrics, 2008) - clustering of correlated predictors to enhance predictive performance.

Properties of LASSO estimators (1)

Function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is said to be pseudo-Lipschitz if there is a numerical constant L such that for all $x, y \in \mathbb{R}^2$,

$$|\varphi(x) - \varphi(y)| \leq L(1 + \|x\|_{\ell_2} + \|y\|_{\ell_2})\|x - y\|_{\ell_2}.$$

Properties of LASSO estimators (1)

Function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is said to be pseudo-Lipschitz if there is a numerical constant L such that for all $x, y \in \mathbb{R}^2$,

$$|\varphi(x) - \varphi(y)| \leq L(1 + \|x\|_{\ell_2} + \|y\|_{\ell_2})\|x - y\|_{\ell_2}.$$

For any $\delta > 0$, $\alpha_{\min} = \alpha_{\min}(\delta)$ is the unique solution to

$$2(1 + \alpha^2)\Phi(-\alpha) - 2\alpha\phi(\alpha) - \delta = 0$$

if $\delta \leq 1$, and 0 otherwise.

Properties of LASSO estimators (2)

Theorem (Theorem 1.5 of Bayatti and Montanari, 2012)

Consider the linear model with i.i.d. $\mathcal{N}(0,1)$ errors in which X is an $n \times p$ matrix with i.i.d. $\mathcal{N}(0,1/n)$ entries. Suppose that the β_i's are i.i.d. random variables, independent of X, and with positive variance (below, Θ is a random variable distributed as β_i).

Properties of LASSO estimators (3)

Theorem

Then for any pseudo-Lipschitz function φ, the lasso solution $\hat{\beta}$ to (3) *with fixed λ obeys*

$$\frac{1}{p}\sum_{i=1}^{p}\varphi(\hat{\beta}_i,\beta_i) \longrightarrow \mathbb{E}\varphi(\eta_{\alpha\tau}(\Theta+\tau Z),\Theta), \tag{11}$$

where the convergence holds in probability as $p, n \to \infty$ in such a way that $n/p \to \delta$. Above, $Z \sim \mathcal{N}(0,1)$ independent of Θ, and $\tau > 0, \alpha > \alpha_{\min}(\delta)$ are the unique solutions to

$$\begin{aligned} \tau^2 &= 1 + \frac{1}{\delta}\mathbb{E}\Big(\eta_{\alpha\tau}(\Theta+\tau Z)-\Theta\Big)^2, \\ \lambda &= \Big(1-\frac{1}{\delta}\mathbb{P}(|\Theta+\tau Z| > \alpha\tau)\Big)\alpha\tau. \end{aligned} \tag{12}$$

Multiple testing notions in multiple regression

$\varphi_V(x,y) = 1(x \neq 0)1(y = 0)$, $\varphi_R(x,y) = 1(x \neq 0)$,
$\varphi_F(x,y) = 1(y \neq 0)$
so that the number V of false discoveries is equal to

$$V = \sum_i \varphi_V(\hat{\beta}_i, \beta_i),$$

the number R of discoveries is equal to

$$R = \sum_i \varphi_R(\hat{\beta}_i, \beta_i).$$

and the number F of true regressors is equal to

$$F = \sum_i \varphi_F(\hat{\beta}_i, \beta_i).$$

Multiple testing notions in multiple regression

$\varphi_V(x,y) = 1(x \neq 0)1(y = 0)$, $\varphi_R(x,y) = 1(x \neq 0)$,
$\varphi_F(x,y) = 1(y \neq 0)$
so that the number V of false discoveries is equal to

$$V = \sum_i \varphi_V(\hat{\beta}_i, \beta_i),$$

the number R of discoveries is equal to

$$R = \sum_i \varphi_R(\hat{\beta}_i, \beta_i).$$

and the number F of true regressors is equal to

$$F = \sum_i \varphi_F(\hat{\beta}_i, \beta_i).$$

$$FDP \equiv V/R$$

Multiple testing notions in multiple regression

$\varphi_V(x, y) = 1(x \neq 0)1(y = 0)$, $\varphi_R(x, y) = 1(x \neq 0)$,
$\varphi_F(x, y) = 1(y \neq 0)$
so that the number V of false discoveries is equal to

$$V = \sum_i \varphi_V(\hat{\beta}_i, \beta_i),$$

the number R of discoveries is equal to

$$R = \sum_i \varphi_R(\hat{\beta}_i, \beta_i).$$

and the number F of true regressors is equal to

$$F = \sum_i \varphi_F(\hat{\beta}_i, \beta_i).$$

$$FDP \equiv V/R$$

FDR for LASSO, Su, Bogdan, Candès (2014)

Theorem

Consider the regression model where X is an $n \times p$ Gaussian design matrix with iid entries following $N(0, \frac{1}{n})$, β_i's are iid random variables with bounded second moment, $z \sim N(0,1)$ and X, β and z are independent. We denote by Θ, a random variable with the same distribution as β_i's. Then it holds that in the limit $p \to \infty$ and $\frac{p}{n} \to \gamma$

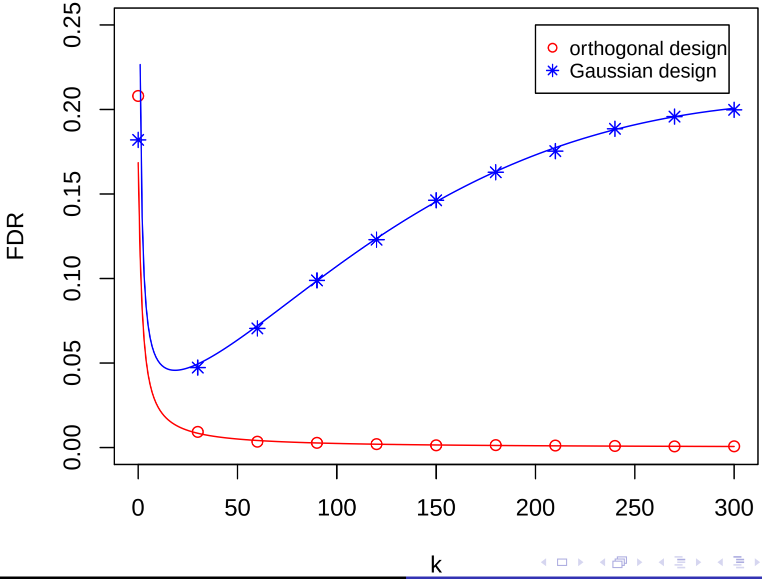
$$FDR \to \frac{2\mathbb{P}(\Theta = 0)\Phi(-\alpha)}{\mathbb{P}(|\Theta + \tau Z| > \alpha\tau)} \; ,$$

$$Power \to \; P \; \mathbb{P}(|\Theta + \tau Z| > \alpha\tau | \Theta \neq 0).$$

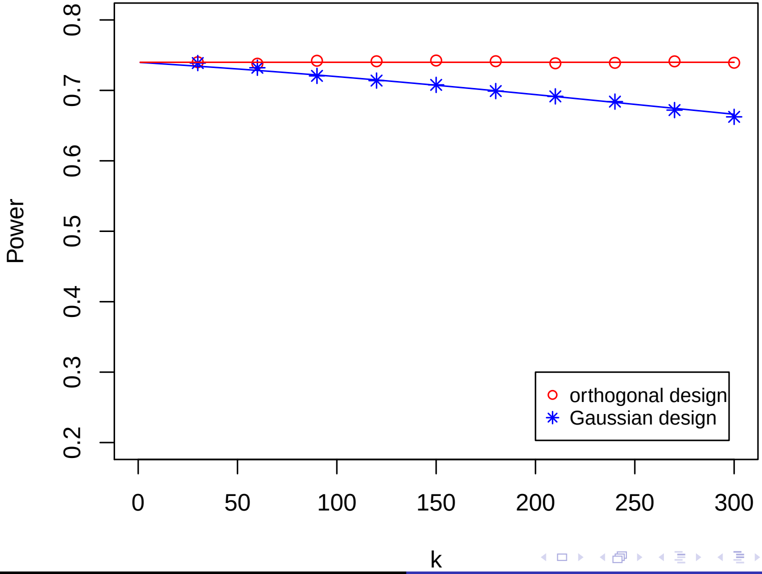
$$\tau^2 = 1 + \gamma \mathbb{E}\Big(\eta_{\alpha\tau}(\Theta + \tau Z) - \Theta\Big)^2$$

$$\lambda = \Big(1 - \gamma \mathbb{P}(|\Theta + \tau Z| > \alpha\tau)\Big)\alpha\tau,$$

FDR - illustration



Power - illustration



minimax FDR (1)

What we believe in

$$\lim_{p,n\to\infty} \inf_{\lambda} \sup_{\beta:||\beta||_{\ell_0}\leq k} \mathrm{FDR}_{\mathrm{lasso}}(\beta,\lambda) = q^\star(\epsilon,\delta), \tag{13}$$

where in the limit, $n/p \to \delta > 0$ and $k/p \to \epsilon > 0$

minimax FDR (1)

What we believe in

$$\lim_{p,n\to\infty} \inf_{\lambda} \sup_{\beta:\|\beta\|_{\ell_0}\le k} \text{FDR}_{\text{lasso}}(\beta,\lambda) = q^\star(\epsilon,\delta), \tag{13}$$

where in the limit, $n/p \to \delta > 0$ and $k/p \to \epsilon > 0$

What we have

Theorem

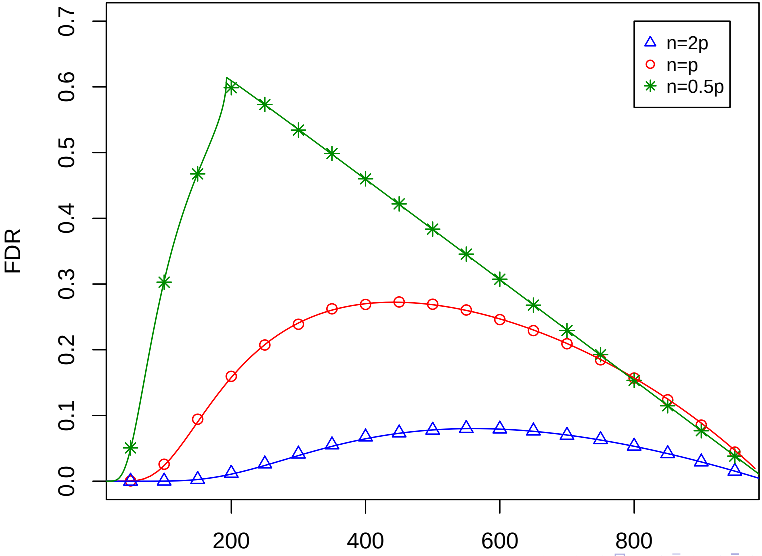
For any $c > 0$ and $\epsilon < \epsilon_\gamma$ if $\gamma > 1$, we have

$$\inf_{\lambda>0} \sup_{\Theta\in\mathcal{F}_c^\epsilon} \lim_{p\to\infty} FDR(n,p,\Theta,\lambda) = FDR_m(\gamma,\epsilon).$$

$\mathcal{F}_c^\epsilon$ is the family of all distributions satisfying

- $\Theta \neq 0$ with probability ϵ.
- If $\Theta \neq 0$, then $|\Theta| > c$ a.s.
- Θ has finite second moment.

minimax FDR (2), $\lambda = 300$, $M = 300000$



Limit on power (1)

Theorem

Let $\epsilon^\star = \epsilon^\star(\delta)$ denote the point on the transition curve. Let us define a function

$$\gamma^\star(\epsilon,\delta) \triangleq \begin{cases} 1 - \frac{(1-\delta)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}, & \delta < 1 \text{ and } \epsilon > \epsilon^\star \\ 1, & \text{otherwise.} \end{cases}$$

It holds

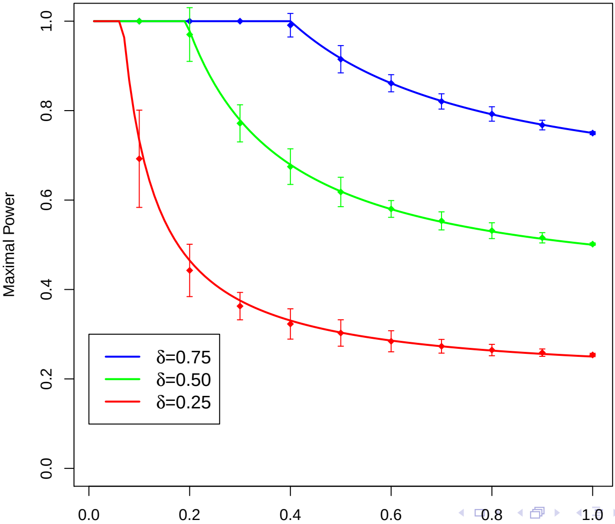
(a)

$$\lim_{p\to\infty} \sup_{\lambda\in(0,\infty),\pi\in\Omega(\epsilon)} \text{Power}(\lambda,\pi,p,\delta) = \gamma^\star(\epsilon,\delta)$$

(b) *for any constants $\lambda_0 > 0$ and $\nu > 0$, with probability tending to one,*

$$\sup_{\lambda_0<\lambda<\infty} < \gamma^\star(\epsilon,\delta) + \nu.$$

Limit on power (2)



Trade off between Power and FDR

Theorem

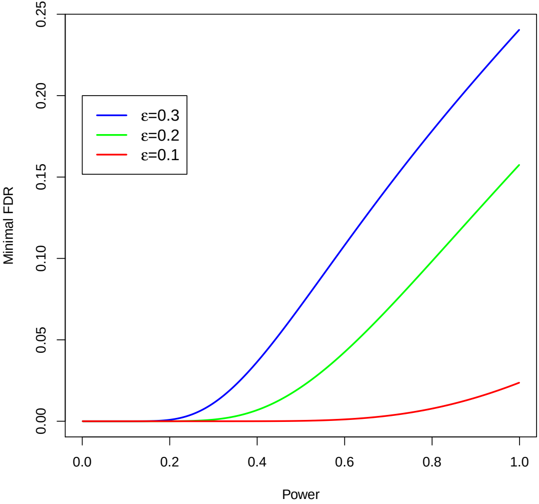
Given Power larger than or equal to $\beta \in (0, \min\{1, \delta\})$, *the minimum of FDR is given as*

$$\mathrm{FDR}_{\min} = \frac{2(1-\epsilon)\Phi(-\alpha_{\max})}{2(1-\epsilon)\Phi(-\alpha_{\max}) + \epsilon\beta}.$$

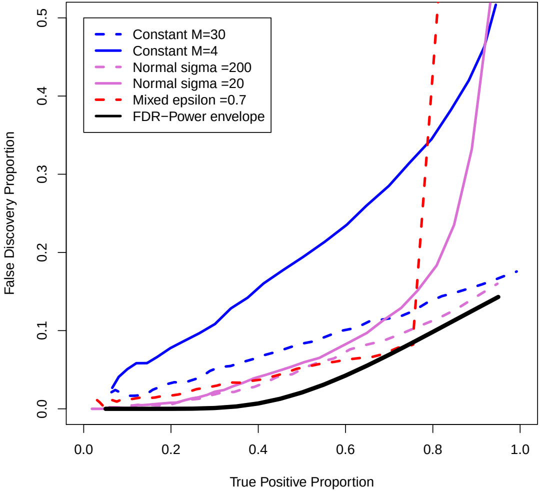
$$\frac{(1-\epsilon)\left[2(1+\alpha_{\max}^2)\Phi(-\alpha_{\max}) - 2\alpha_{\max}\phi(\alpha_{\max})\right] + \epsilon(1+\alpha_{\max}^2) - \delta}{\epsilon\left[(1+\alpha_{\max}^2)(1-2\Phi(-\alpha_{\max})) + 2\alpha_{\max}\phi(\alpha_{\max})\right]} =$$

$$\frac{1-\beta}{1-2\Phi(-\alpha_{\max})}.$$

FDP-TPP tradeoff



FDP-TPP tradeoff



FDP-TPP tradeoff