Large Scale Inference when Signals are Rare and Weak

Instructors: Xin Tong (National University of Singapore)
Wanjie Wang (National University of Singapore)

Tutorial on High-dimensional Data Analysis

IMS Singapore, Feb. 5, 2018

Big Data everywhere

- Microarray, GWAS, DNA sequencing, ...
- Image processing
- Tick-by-tick financial data
- Large-scale social networks
- ...

White House "Big Data Research and Development Initiative", March 2012

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges.

To launch the initiative, six Federal departments and agencies today announced more than $200 million in new commitments that, together, promise to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

Sparsity: a direct result of "large p"

Exact meanings vary, but are similar:

- parameter sparsity (e.g., normal means model)
- transform coefficient sparsity (e.g., wavelet)
- graph sparsity (e.g., Graphical model)
- sparsity in the eigen-space (e.g., spike model)

Exploiting sparsity has become a major strategy for analyzing Big Data

Weak signals: a direct result of "small n"

e.g. Why most published research findings are false (Ioannidis, 2001, PLoS Med)

$$(\text{Signal strength})^2 \quad \propto \quad n \quad \propto \quad \$ \text{ or manpower}$$

- Example: GWAS, Rare Variation Association Study (RVAS), Gravitational wave.
- An important but largely neglected notion
- Partially explains
  - irreproducibility of many published works
  - many problems remain unsolved by increasing p
  - theoretical optimism vs. practical pessimism

Rare/Weak signals

- **Rare**. only a small fraction of returns contain tradable signals, others are merely noise
- **Weak**. signals are individually weak

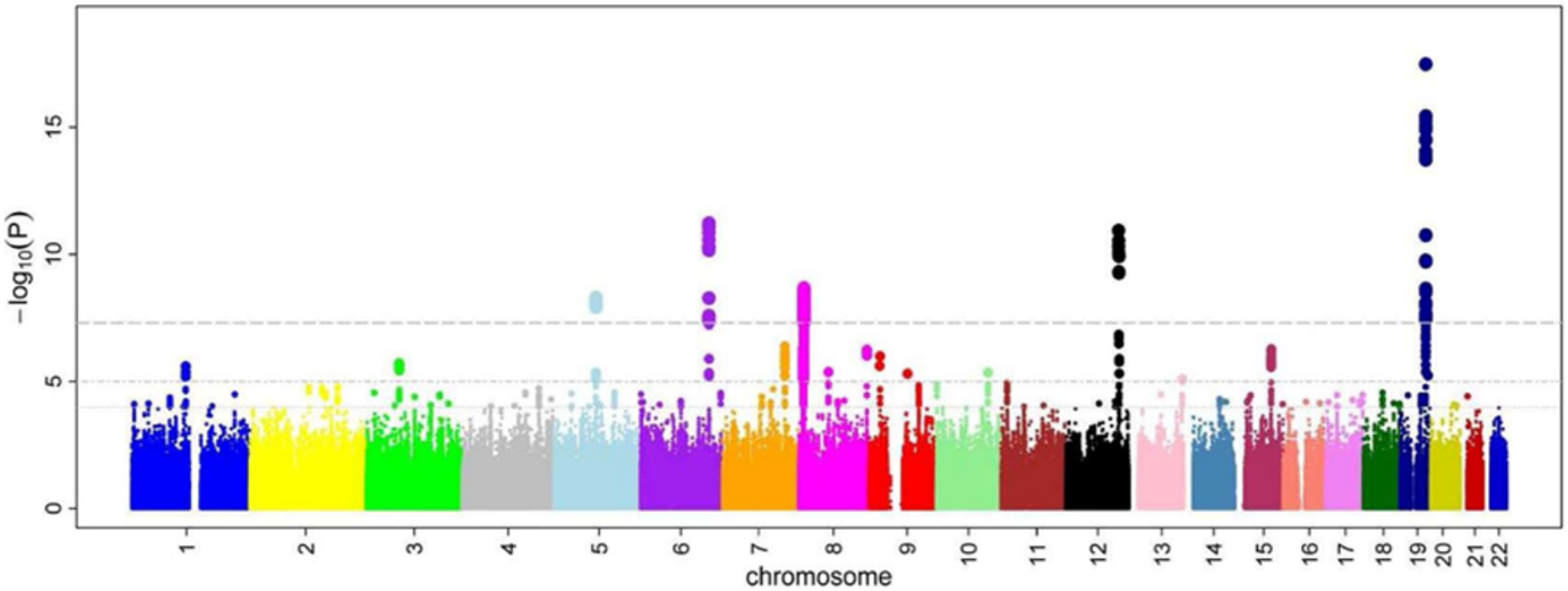


Figure is downloaded from Wikipedia

Normal means model

$$X = \mu + z, \qquad z \sim N(0, \Sigma), \qquad \Sigma \in \mathbb{R}^{p,p}$$

- There are p experiments (p is presumably large)
- X_j: z-score from experiment j
- μ_j: (unknown) true effect in experiment j
  A nonzero μ_j means a true discovery ("signal")
- **Rare**: only a small fraction of μ_j's are nonzero
- **Weak**: each μ_j is small in magnitude

Four interconnected problems

- **Global testing**. Are there any signals at all?
- **Classification**. Using in-sample to select a few useful features for out-sample prediction
- **Clustering**. Classification when no training sample is available
- **Signal recovery**. Identify signals individually

Overview I. Higher Criticism for global testing

- Basic ideas and phase transition:

  Donoho D, Jin J (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* **32**(3) 962-994.

- Review on ideas, methods, and applications:

  Donoho D, Jin J (2015) Higher Criticism for Large-Scale Inference: especially for Rare and Weak effects. *Statistical Science* **30**(1) 1-25 (Invited Review Paper).

- Review on methods and theory, especially phase transition:

  Jin J, Ke T Z (2015) Rare and Weak effects in Large-Scale Inference: methods and phase diagrams. *Statistica Sinica* **26** 1-34 (Invited Review Paper).

Overview, II. Classification and clustering

- Part 2. Fisher's LDA for Classification
  - HC for threshold choice in feature selection
  - estimating $\Omega = \Sigma^{-1}$
- Part 3. clustering by Influential-Features PCA
  - PCA applied to a few carefully selected features
  - HC for threshold choice in feature selection
  - use Efron's empirical null to minimize the gap between theory and data analysis
- Part 4. Graphlet screening for variable selection
  - connected to Innovated HC
  - exploits "local graphical" structure for screening

Part I: Global Testing

Global testing

Given a large number of

- summarizing t-scores or z-scores
- or summarizing P-values
- or regression coefficients

Test

$$H_0 : \quad \text{contains no signals}$$

vs.

$$H_1 : \quad \text{contains Rare/Weak signals}$$

Normal means model: $X \sim N(\mu, \Sigma)$

$$H_0 : \mu = 0, \qquad H_1 : \text{some } \mu_j\text{'s are nonzero}$$

Simple methods and their limitations

Chi-square test:

$$T = \sum_{j=1}^{p} X_j^2$$

- Works when many X_j's contain signals
- Low SNR when signals are rare, because most X_j's are merely noise

Minimum P-value (minP) test:†

$$T = \max_{1 \le j \le p} |X_j|$$

- Works for rare signals
- Sub-optimal for Rare and *very* Weak signals

† *It is equivalent to using the smallest P-value as the test statistic*

1.1. Higher Criticism (HC)

Higher Criticism in Tukey's Lecture Notes

John W. Tukey (1915-2002)

t21(exT4-1

1976 Statistics 411

T31(exT21(exT4))

THE HIGHER CRITICISM AND KINDS OF ERROR RATES

Once we deal with parallel estimates -- we will take parallel centerings for our prototype, but the same questions arise wherever there is parallelism -- we have problems concerning significance, confidence, etc. These problems can have more than one resolution, but the more unhappy resolutions (in terms of discovering less) are often those that seem better justified when we consider things carefully.

T4A. The simple higher criticism

There is always the story about the young psychologist -- once he was said to go to a pilot training school, but today he is perhaps more likely to be said to go to the management training school of some large business or industrial organization. Wherever he goes he takes 250 different psychological tests (which we will hereafter call psychological procedures, to avoid two uses of the word "tests").

-- and he comes back full of joy because 12 -- count them, twelve -- of his procedures are significant (at 5%) predictors of performance (as a pilot, manager, or what have you). He is ready to do good things!

Tukey's story

- A young scientist administers 250 uncorrelated tests, where 11 were significant at the 5% level.
- Question: Is this surprising (e.g., any of the test is truly significant)?
- Answer: No, we expect

$$250 \times 5\% = 12.5 \quad \text{significances at 5\% level}$$

Tukey's Higher Criticism

Abstraction: given P-values $\pi_1, \pi_2, \ldots, \pi_p$, test

$H_0^{(p)}$: all p (individual) tests contain nothing but noise

$H_1^{(p)}$: some of the (individual) tests contain signals

$$HC_{p,.05} = \sqrt{p}\left[\frac{(\text{Fraction Significant at } .05) - .05}{\sqrt{.05 \times .95}}\right]$$

- Reject global null $H_0^{(p)} \iff HC_{p,.05} \geq 2$ (say)
- For Tukey's story:

$$HC_{p,.05} = \frac{11 - 12.5}{\sqrt{250 \times .05 \times 0.95}} = -.43 \implies \text{Accept } H_0^{(p)}$$

Donoho&Jin's Higher Criticism

Donoho and Jin (2004)

$$HC_p^* = \max_{0 \le \alpha \le \frac{1}{2}} \left\{ \sqrt{p} \left[\frac{(\text{Fraction Significant at } \alpha) - \alpha}{\sqrt{\alpha(1-\alpha)}} \right] \right\}$$

- A way to combine sparse signals
- Sensitive to "unusually many" moderately strong/weak signals
- Similar to Anderson-Darling statistic (Anderson and Darling (1952); on goodness-of-fit) but different

Implementation of Higher Criticism (HC)

We only need P-values to implement HC

- Sort P-values:

$$\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$$

- Calculate the k-th HC-score

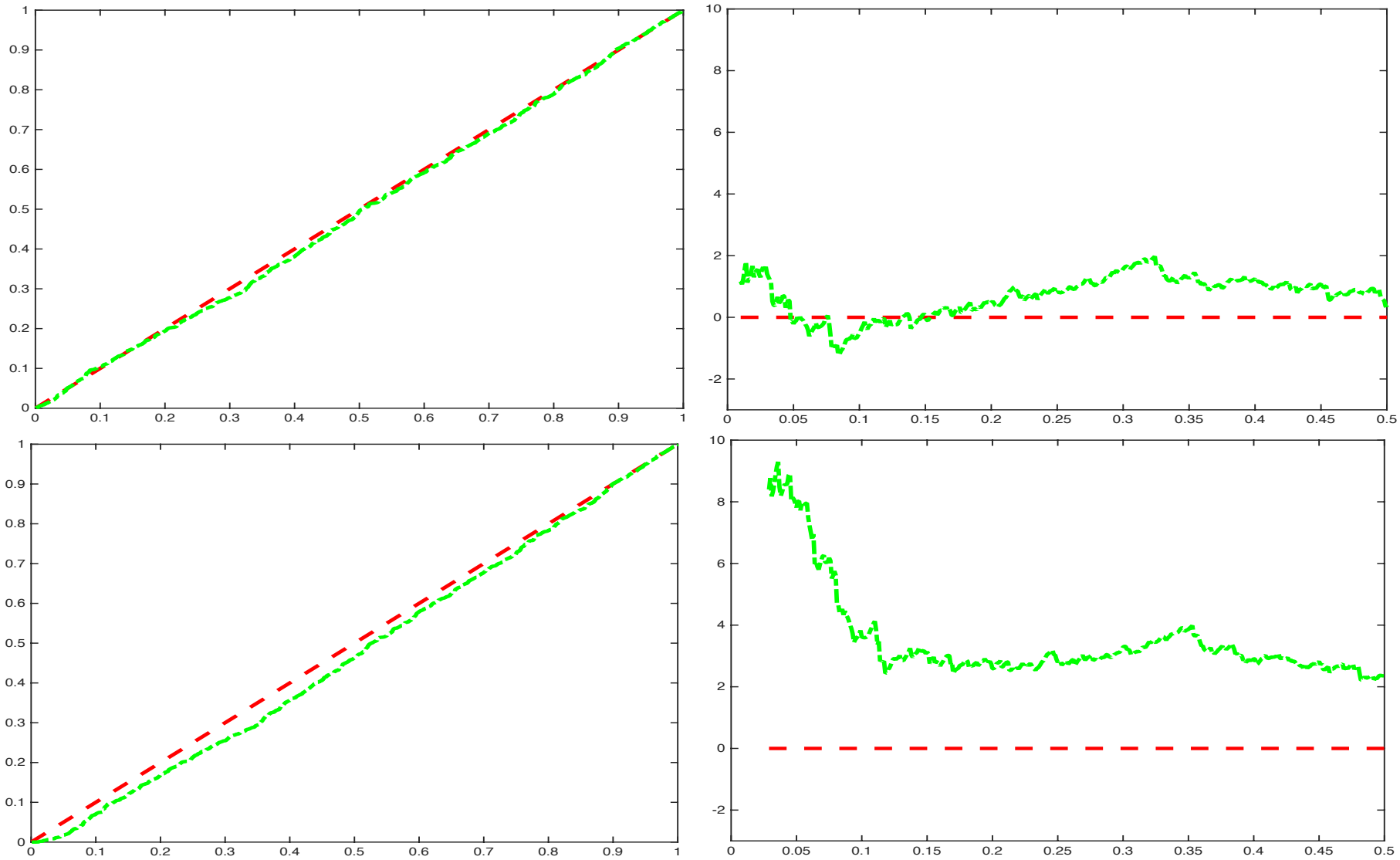
$$HC_{p,k} = \sqrt{p}\left[\frac{k/p - \pi_{(k)}}{\sqrt{\pi_{(k)}(1-\pi_{(k)})}}\right],$$

- Take maximum:

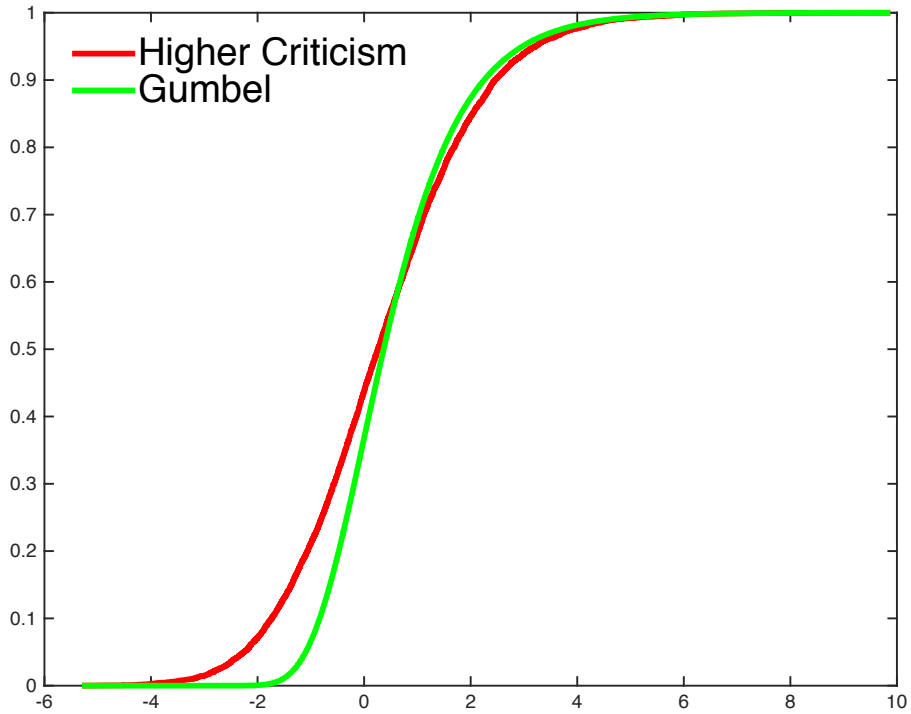
$$HC_p^+ = \max_{\{1\le k\le p/2,\ \pi_{(k)}\ge \log(p)/p\}} \{HC_{p,k}\}$$

Illustration

P-values (left two panels) and $HC_{p,K}$ (right two panels) under the null (top two panels) and the alternative (bottom two panels)



Limiting distribution of HC under null



CDF: HC (simulated) vs. Gumbel

- $b_p = \sqrt{2 \log \log(p)}$
- $c_p = \sqrt{2 \log \log(p)} + \frac{1}{2}[\log \log \log(p)] - \log(4\pi)]$
- $b_p HC_p^+ - c_p \Longrightarrow$ Standard Gumbel dist. (CDF: $e^{-e^{-x}}$)

Critical value

Let $h(p, \alpha)$ be the critical value that satisfies

$$P(HC_p^+ \geq h(p, \alpha)) = \alpha.$$

A good approximation of $h(p, \alpha)$ is

$$h_{Gum}(p, \alpha) = \frac{1}{b_p}\left[c_p - \log\log\left(\frac{1}{1-\alpha}\right)\right] \approx \sqrt{2\log\log(p)}$$

p	10^3	5×10^3	2.5×10^4	1.25×10^5
$\alpha = .05$	3.17 (3.00)	3.22 (3.08)	3.26 (3.14)	3.30 (3.19)
$\alpha = .01$	3.95 (3.83)	3.97 (3.87)	3.96 (3.90)	3.99 (3.93)
$\alpha = .005$	4.29 (4.18)	4.28 (4.20)	4.26 (4.22)	4.28 (4.24)
$\alpha = .001$	5.03 (5.00)	5.02 (4.98)	4.98 (4.97)	4.98 (4.97)

1.2. Rare/Weak signal model, phase transitions

Rare/Weak signal model

$$X = \mu + z, \qquad z \sim N(0, \Sigma), \qquad \Sigma \in \mathbb{R}^{p,p}$$

$$H_0^{(p)} : \ \mu = 0 \qquad \textit{vs.} \qquad H_1^{(p)} : \ \mu_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}$$

- $\epsilon_p = p^{-\beta}$, $1/2 < \beta < 1$ (very rare)
- $\tau_p = \sqrt{2r\log(p)}$, $0 < r < 1$ (weak)
- $\Omega = \Sigma^{-1}$ sparse and known

Subtlety of the problem:

$$\sqrt{p}\bar{X} \sim N(p^{\frac{1}{2}-\beta}\tau_p, 1), \qquad p^{\frac{1}{2}-\beta}\tau_p \approx 0 \text{ since } \beta > 1/2$$

$$\max_j |X_j| \sim \sqrt{2\log(p)} \ \text{ under null}, \qquad \text{but } \ \tau_p < \sqrt{2\log(p)}$$

Detecting RW signals by HC

To apply HC, we need to choose the **level** α:

- obtain P-values by

$$\pi_k = P(N(0,1) \geq X_k), \qquad k = 1, 2, \ldots, p$$

- obtain HC_p^+ by three previous steps
- reject null if

$$HC_p^+ \geq h(p, \alpha_p), \qquad \alpha_p: \quad \text{Type I error},$$

for an $\alpha_p \to 0$ slowly enough (e.g., $\alpha_p = \frac{1}{\log(p)}$)

Phase transition

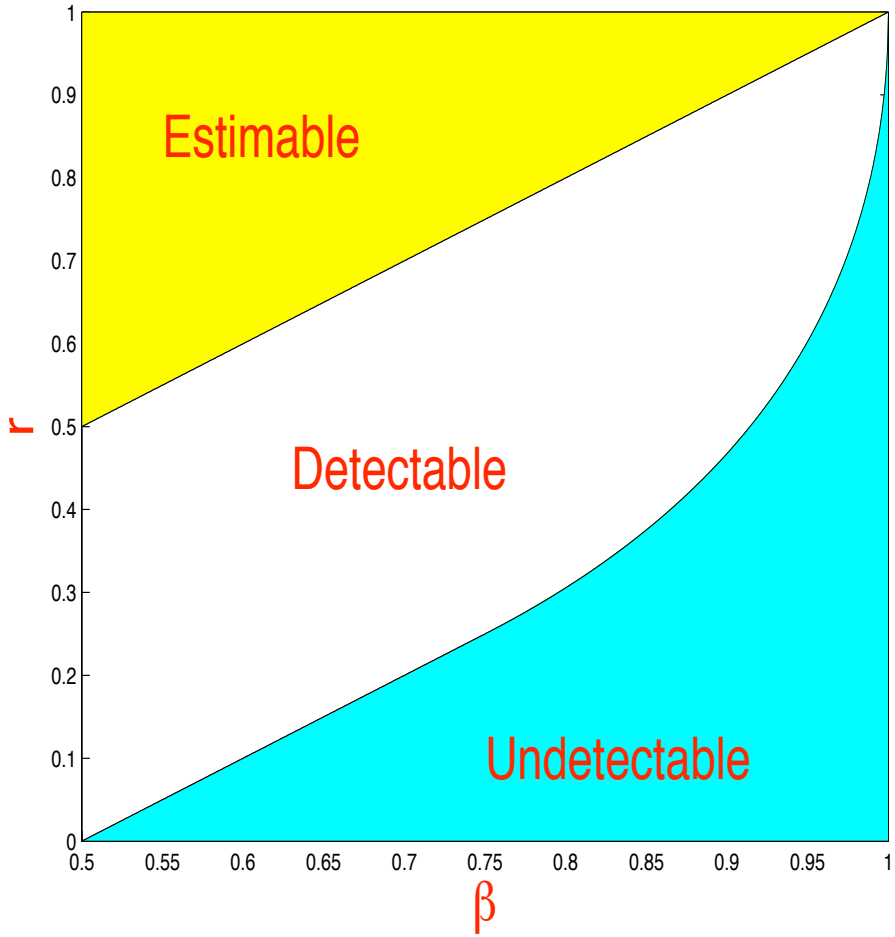
Define *standard phase function*

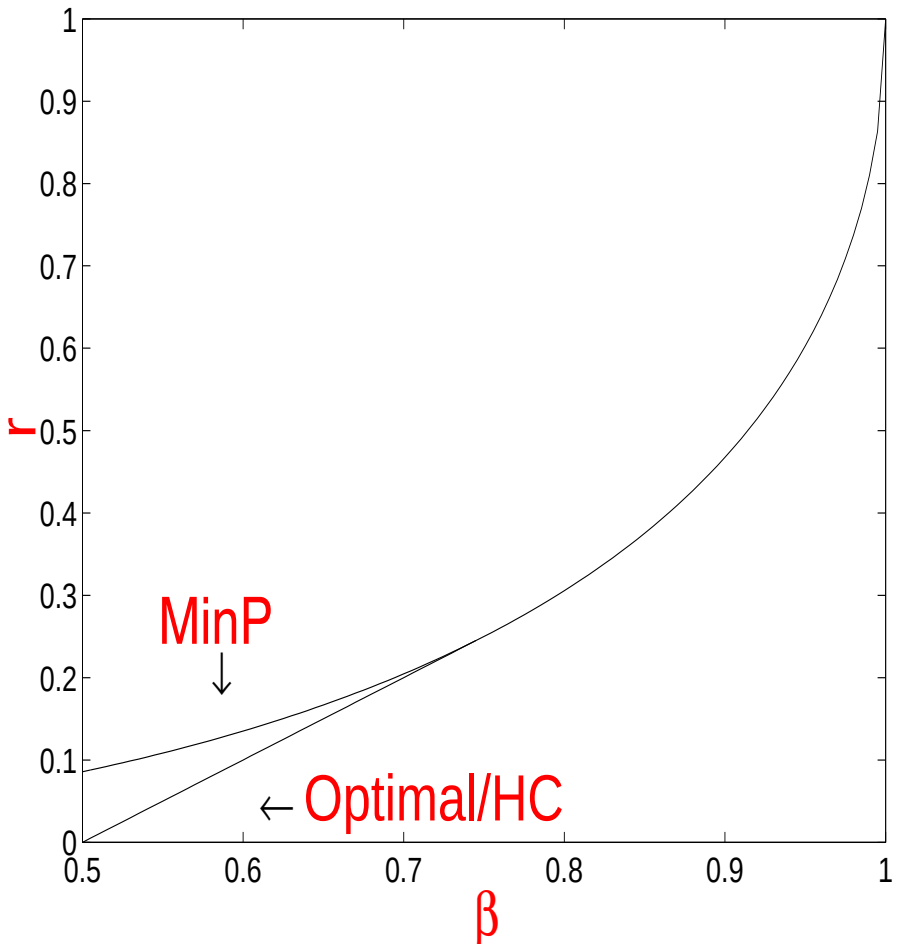
$$\rho(\beta) = \begin{cases} 0, & 0 < \beta < 1/2 \\ \beta - \frac{1}{2}, & \frac{1}{2} < \beta < \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2, & \frac{3}{4} < \beta < 1 \end{cases}$$

Theorem 1. As $p \to \infty$,

- If $r < \rho(\beta)$, for any test, (Type I + Type II) errors $\to 1$
- If $r > \rho(\beta)$, both Type I and Type II errors of HC $\to 0$.
- Such a result does not hold for many tests ($\bar{X}$, χ^2, MinP)

Phase diagram (signal detection, $\Sigma = I_p$)





Signal recovery (variable selection)

- Goal: identify which μ_j's are nonzero
- Hamming loss

$$Hamm(\hat{\mu}, \mu) = \sum_{j=1}^{p} 1\{\text{sign}(\hat{\mu}_j) \neq \text{sign}(\mu_j)\}$$

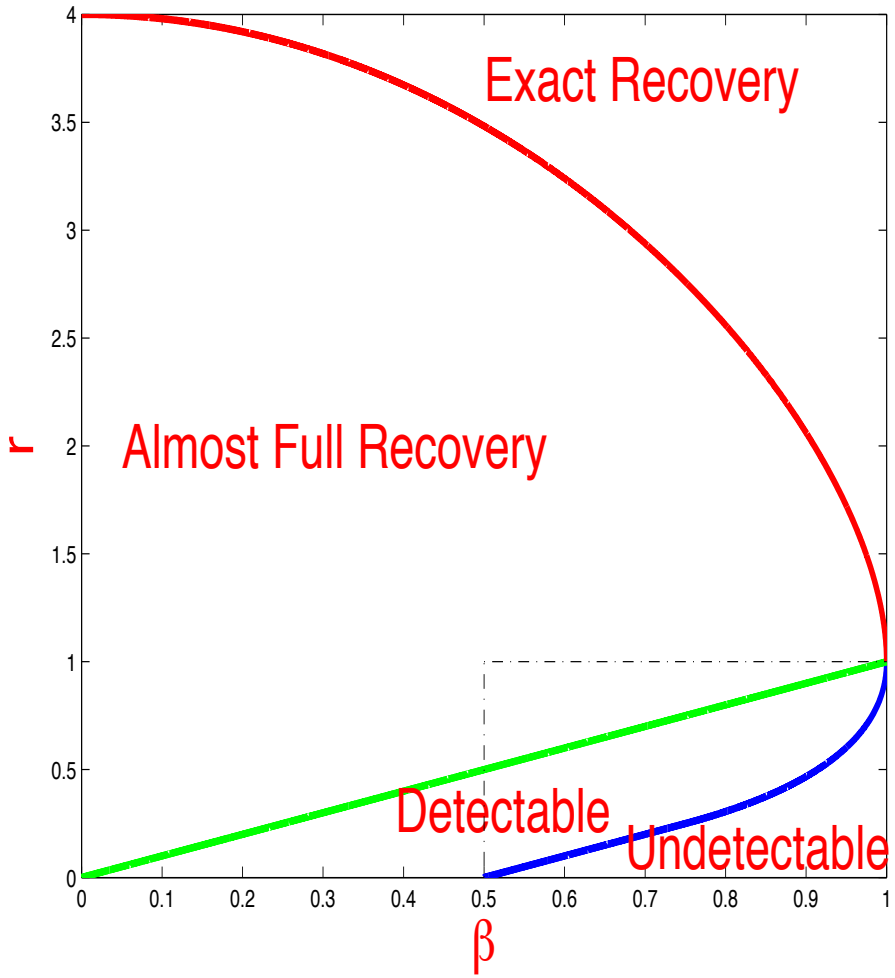
- Three regions:
  - Exactly recoverable: $Hamm(\hat{\mu}, \mu) < 1$
  - Almost fully recoverable: $1 \leq Hamm(\hat{\mu}, \mu) \ll p\epsilon_p$
  - No recoverable: $Hamm(\hat{\mu}, \mu) \asymp p\epsilon_p$
- A variable selector $\hat{\mu}$ (for $\Sigma = I_p$)

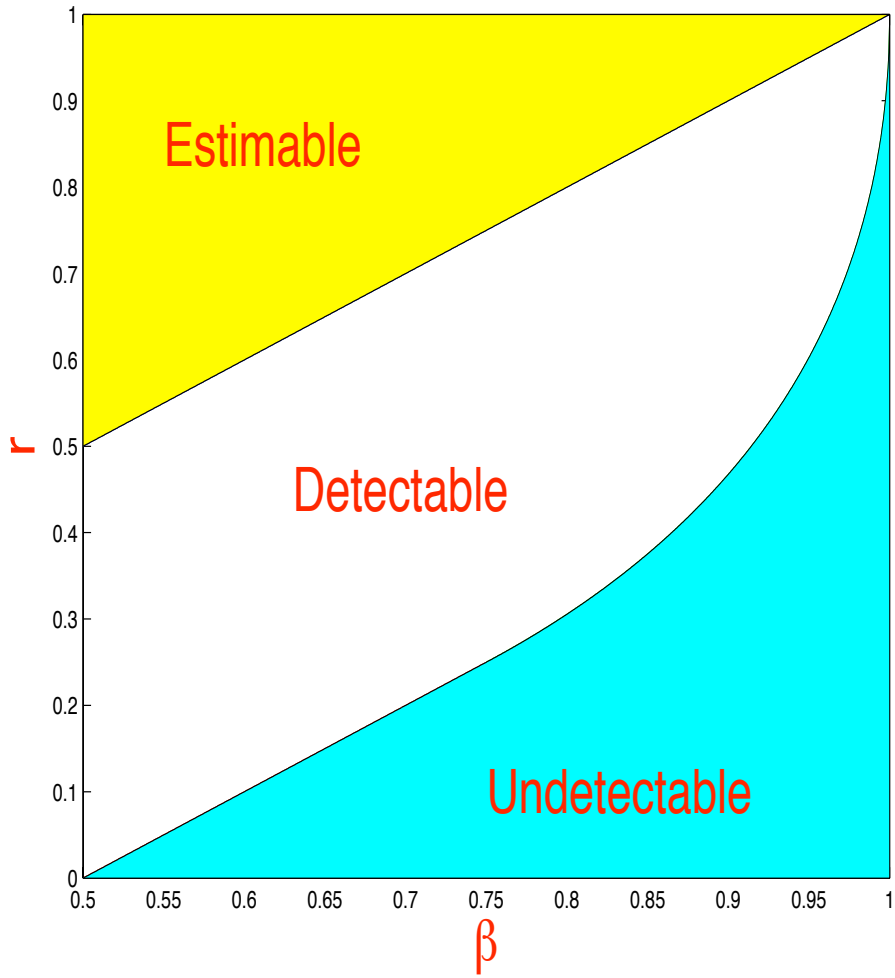
$$\hat{\mu}_j = X_j 1\{|X_j| > \sqrt{2t\log(p)}\}$$

† *Here* $\text{sign}(\cdot)$ *takes values in* $\{-1, 0, 1\}$

Phase diagram: a more complete story

Extended to more complicate cases (e.g., $\Sigma \neq I_p$, regression models); see the review paper by Jin and Ke (2015)





1.3. Innovated Higher Criticism

HC for detecting RW signals ($\Sigma \neq I_p$)

$$y = \mu + z, \qquad z \sim N(0, \Sigma), \qquad \Sigma \in \mathbb{R}^{p,p}$$

We only need P-values to implement HC

- Transform the data: For a $p \times p$ matrix A,

$$y \mapsto \tilde{y} \equiv Ay \sim N(A\mu, A\Sigma A')$$

- Compute the marginal P-values

$$\tilde{y}_j \sim N(0, (A\Sigma A')_{jj}) \text{ under the null} \quad \Longrightarrow \quad \pi_j$$

- Plug $\pi_1, \pi_2, ..., \pi_p$ into the HC procedure

Different transforms

- Brute-force:

$$y \sim N(\mu, \Sigma), \qquad \pi_j = P(N(0, \Sigma_{jj}) \geq y_j)$$

- Whitening:

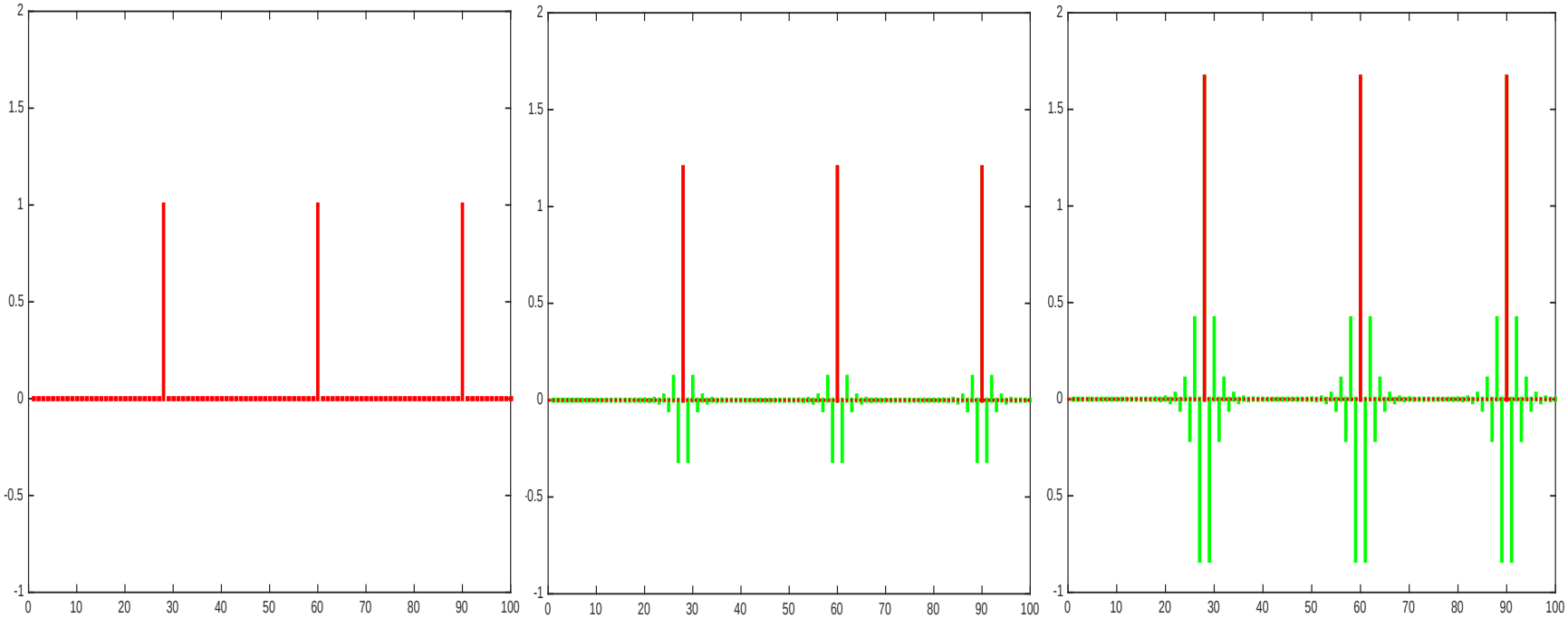
$$\tilde{y} = \Sigma^{-\frac{1}{2}} y \sim N(\Sigma^{-\frac{1}{2}}\mu, I_p), \quad \pi_j = P(N(0,1) \geq \tilde{y}_j)$$

- Innovated Tranform (recall $\Omega = \Sigma^{-1}$):

$$\tilde{y} = \Sigma^{-1} y \sim N(\Omega y, \Omega), \quad \pi_j = P(N(0, \Omega_{jj}) \geq \tilde{y}_j)$$

Comparison of SNR

Innovated Transform $y \mapsto \Sigma^{-1} y$ yields largest SNR at $k_1, k_2, \ldots, k_m$



SNR for Brute-force, Whitening, and Innovated Transform. Σ is tridiagonal, and μ has 3 nonzeros

Why Innovated Transform the right choice

- A is row-wise sparse $\Longrightarrow$ $A\mu$ is still sparse
- A yields a large SNR simultaneously for j

$$SNR_j = \frac{(A\mu)_j}{\sqrt{(A\Sigma A)_{jj}}} \approx \frac{A_{jj}}{\sqrt{(A\Sigma A)_{jj}}}\mu_j$$

a_j': j-th row of A. Let $\xi = \Sigma^{1/2} a_j$ and $\eta = \Sigma^{-1/2} e_j$

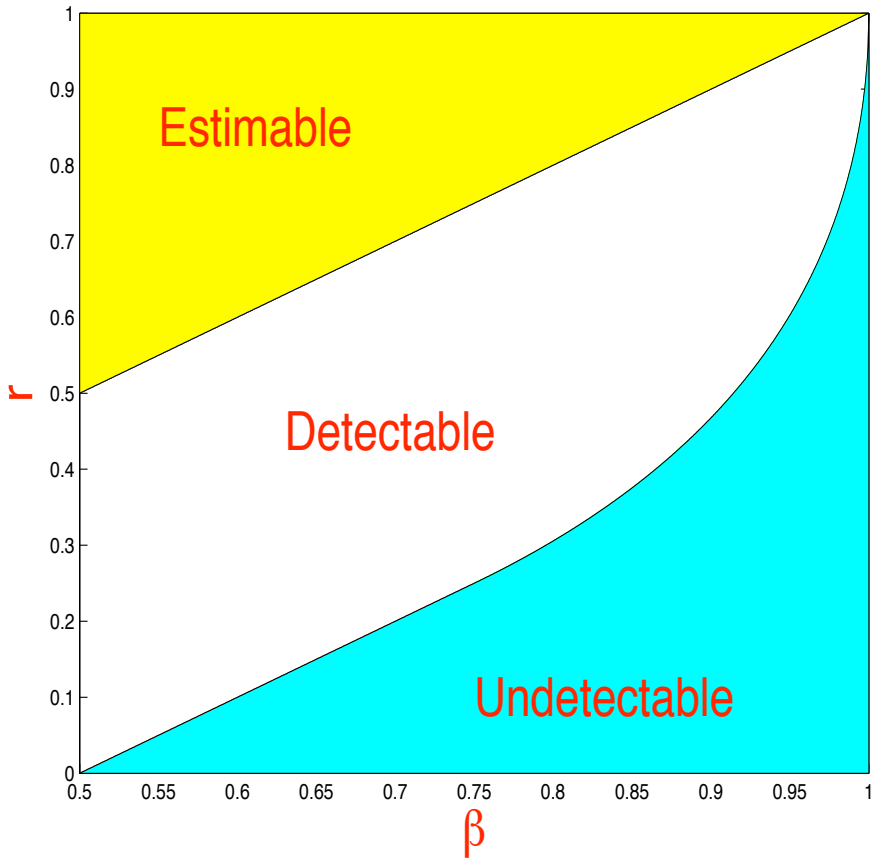
$$\frac{A_{jj}}{\sqrt{(A\Sigma A)_{jj}}} = \frac{a_j' e_j}{\sqrt{a_j' \Sigma a_j}} = \frac{\xi'\eta}{\|\xi\|}, \quad \stackrel{\text{maximize}}{\Longrightarrow} \quad \xi \propto \eta \quad \Longrightarrow \quad a_j \propto \Sigma^{-1} e_j$$

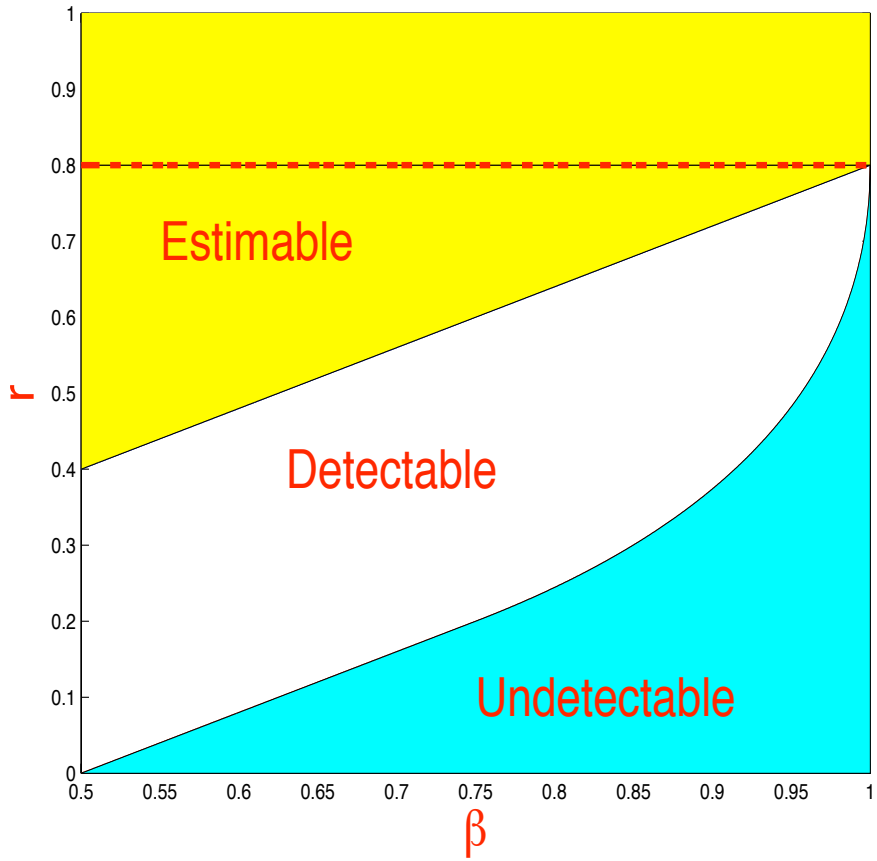
- $A = \Sigma^{-1}$ maximizes the SNR for all j
- Innovated HC: Innovated transform + HC

Correlations: Curse or blessing?

- Noise correlation is a *blessing*: "neighboring entries" contain useful information
- This can be seen by comparing the phase diagrams

  (Left: $\Sigma = I_p$. Right: Σ is Toeplitz with unit diagonals)





Other signal detection settings

HC is ideal for: out of many units or cases, most are null or baseline effects; non-null effects are Rare and individually Weak

- Matrix testing (eigenvalue-based HC)
- Network anomaly detection (degree-based HC)
- Genome-Wide Association Study
- DNA Copy Number Variants (CNV)
- Cosmology and Astronomy
- Disease surveillance
- Anomaly detection
- Estimate the proportion of non-null effects (Kuiper Belt Objects)

Ke (2017); Mukherjee et al (2016); Sabatti et al (2008); Jeng et al (2012); Bennett et al. (2012); McFowland et al (2013); Saligrama and Zhao (2012); Meinshausen and Rice (2004)

Summary of Part I

Higher Criticism for global testing

- only need P-values to use HC
- especially effective RW settings
- flexible, useful in many different settings
- correlation among noise is a blessing
- Innovated Transform is marginal regression in disguise and maximize SNR at $k_1, k_2, \ldots, k_m$
- HC achieves the optimal phase diagrams

Reference

- For HC basics, overview, and applications, see

  Donoho D, Jin J (2015) Higher Criticism for Large-Scale Inference: especially for Rare and Weak effects. *Statistical Science* **30**(1) 1-25 (Invited Review Paper).

- For phase diagrams, innovated transform, and connections to regression models, see

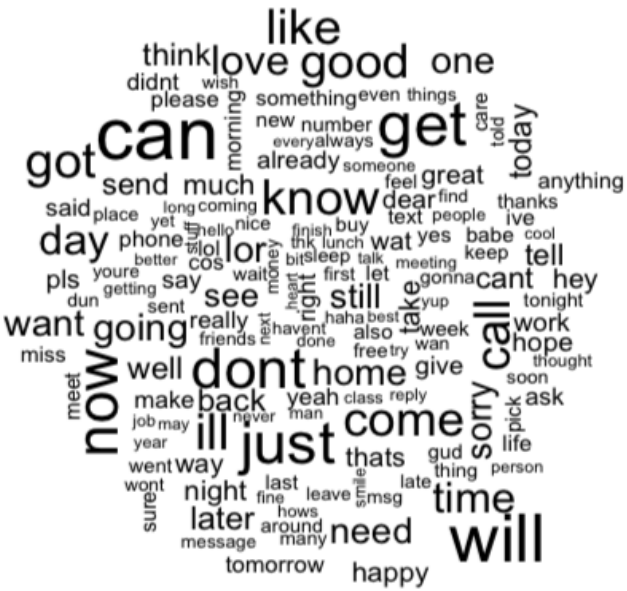
  Jin J, Ke T Z (2015) Rare and Weak effects in Large-Scale Inference: methods and phase diagrams. *Statistica Sinica* **26** 1-34 (Invited Review Paper).

Part II: Classification

Two-class classification

- n training samples (X_i, Y_i)
  - $X_i \sim N(Y_i \cdot \mu, \ \Sigma)$: feature vectors in $\mathbf{R^p}$
  - $Y_i = \pm 1$: class labels
- **Goal**. given test sample vector X, predict class label $Y = 1$ or $Y = -1$

Good email



spam

(Pictures here are downloaded online)

Fisher's Linear Discriminant Analy. (LDA)

$$L(X) = \sum_{j=1}^{p} w(j) \cdot X(j)$$

- $w(j)$: feature weights determined by (X_i, Y_i)
- Classify $Y = \begin{cases} 1, & L(X) > 0 \\ -1, & L(X) < 0 \end{cases}$
- Optimal weights: $w \propto \Omega\mu$, approachable when $n \gg p$, where

$$\Omega = \Sigma^{-1}: \quad \text{precision matrix}$$

Rare/Weak setting

$X_i \overset{iid}{\sim} N(\pm\mu, \Sigma)$, nonzero μ_j's are Rare/Weak

- **Estimating** Ω. Given *iid* samples from $N(0, \Sigma)$,[†] how to estimate Ω
- **Classification**. How to classify, and especially, how to estimate μ, if signals are Rare/Weak and Ω is sparse & known

[†] *In our setting,* $X_i \sim N(\pm\mu, \Sigma)$, *but Rare/Weak nonzero means are thought of having negligible effects in estimating* Ω

Related work

- Case 1. $\Omega = I_p$ (*Donoho and Jin (2008, 2009), Jin (2009)*)
  - Feature selection, threshold choice by HC Thresholding (**HCT**)
  - Classification limits, phase transition
- Case 2. Ω is K-sparse and known (*Fan, Jin, Yao (2013)*)
  - Combine Innovated Transform with HCT
  - A useful lemma in graph theory (TBA)
- Case 3. Ω is sparse and unknown (*Huang, Jin, Yao (2016)*)
  - Estimate Ω by Partial Correlation Screening (**PCS**)
  - Classify by **HCT-PCS**

2.1. The HCT Classifier

Estimating μ when Ω is known & sparse

- Training samples: $X_i \sim N(Y_i \cdot \mu,\ \Sigma),\ Y_i = \pm 1$
- Summarizing Z-scores:

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \cdot X_i \quad \sim \quad N(\sqrt{n} \cdot \mu,\ \Sigma)$$

- Boost SNR by Innovated Transform:

$$\text{Innovated } Z\text{-scores}: \qquad \widetilde{Z} = \Omega Z \ \sim \ N(\sqrt{n} \cdot \Omega\mu,\ \Omega)$$

- Clipping thresholding (**marginal reg. in disguise**):

$$\hat{\mu}_t(j) = \begin{cases} \operatorname{sign}(\widetilde{Z}(j)), & |\widetilde{Z}(j)| \geq t, \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, 2, \ldots, p$$

Question.

- How to set the threshold t?

Threshold choice

- Cross validation (CV)
  - Each CV wastes a valuable fraction of samples
  - Time consuming
  - Comparably unstable
- Control feature-wise False Discovery Rate
  - It needs to set the FDR level q (it becomes another tuning parameter, unclear how to set)

Proposal:

- threshold choice by Higher Criticism (**HC**)
- optimality of this choice

Higher Criticism Threshold (HCT)

Donoho and Jin (2008)

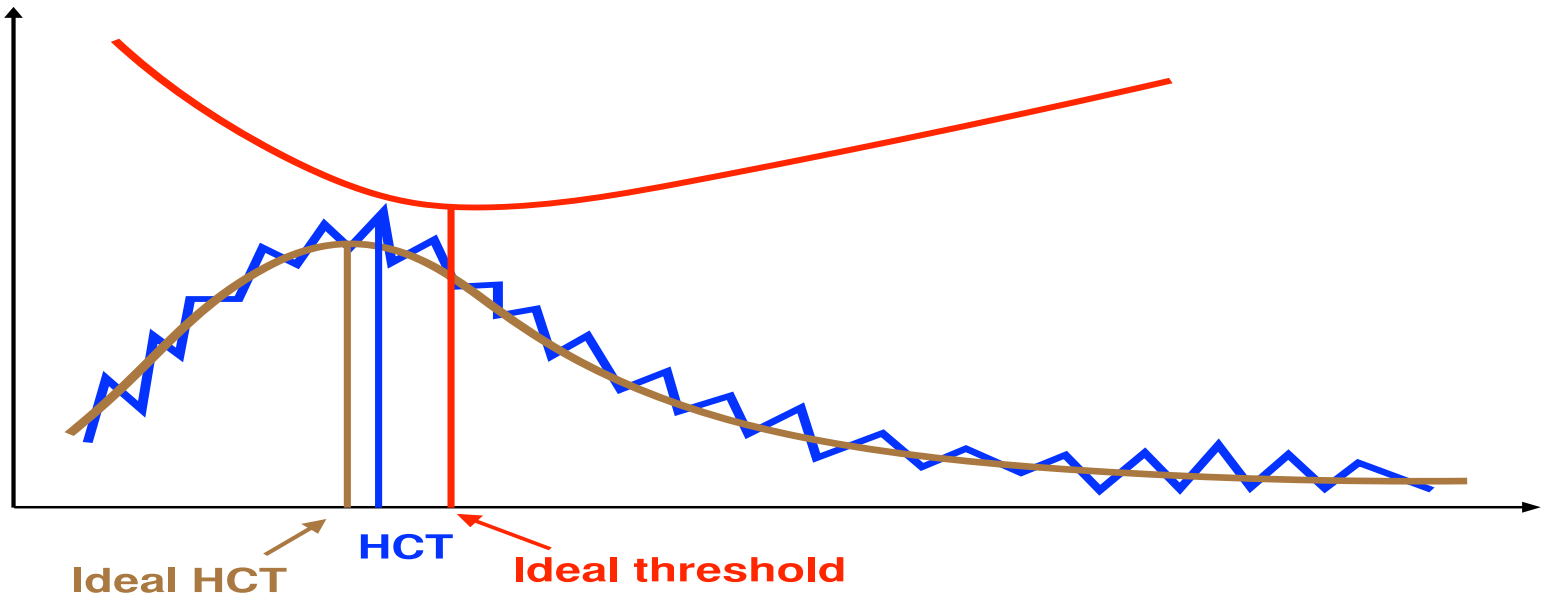
Innovated Z-scores: $\widetilde{Z} \sim N(\sqrt{n}\Omega\mu, \Omega)$

1. Convert to P-values: $\pi_j = P\{|N(0,1)| > \frac{|\widetilde{Z}(j)|}{\sqrt{\Omega(j,j)}}\}$
2. Sort: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$
3. HC-scores:
$$HC_{p,k} = \sqrt{p}\left[\frac{k/p - \pi_{(k)}}{\sqrt{(k/p)(1-k/p)}}\right]$$
4. HC-threshold (HCT):
$$t_p^{HC} = \hat{k}\text{-th largest } Z\text{-score (in absolute value)},$$
$$\text{where } \hat{k} = \text{argmax}_{\{1 \le k \le \frac{p}{10}\}}\{HC_{p,k}\}$$

Comparison of two versions of HC

	Orthodox HC	New HC (HCT)
Objective	Global testing	Threshold choice for feature selection
$HC_{p,k}$	$\sqrt{p}\frac{k/p-\pi_{(k)}}{\sqrt{\pi_{(k)}(1-\pi_{(k)})}}$	$\sqrt{p}\frac{k/p-\pi_{(k)}}{\sqrt{(k/p)(1-k/p)}}$
Statistics	$\max_{\{1\leq k\leq \frac{p}{2}, \pi_{(k)}\geq \frac{\log(p)}{p}\}}\{HC_{p,k}\}$	$\hat{k}=\operatorname{argmax}_{\{1\leq k\leq \frac{p}{10}\}}\{HC_{p,k}\}$

Rationale of HCT



- Red line: true classification error, minimized at *Ideal Threshold* (unknown)
- Blue line: data-driven HC scores, maximized at *HCT*
- Brown line: population counterpart of HC scores, maximized at *Ideal HCT*

In theory, HCT $\approx$ Ideal HCT $\approx$ Ideal Threshold

(Figure is only an illustration of ideas, not from simulations)

Estimating Ω

$$X_i \overset{iid}{\sim} N(0, \Sigma), \qquad \Omega = \Sigma^{-1} \text{ is sparse}$$

- Sample cov. matrix: $\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} X_i X_i'$
- The glasso: estimating Ω by minimizing

$$\log(det(\Omega)) - trace(\hat{\Sigma}\Omega) - \lambda \|\Omega\|_1, \quad \Omega \succ 0$$

- Penalization methods: estimating each row of Ω by a penalized regression *(Dettling and Buhlmann (2003); Cai, Liu, and Luo (2011))*
- PCS: estimating each row of Ω by a Screen-and-Clean method *(Huang, Jin, and Yao (2016))*

Comparison with popular classifiers

A synthetic Rare/Weak setting:

$X_i \overset{iid}{\sim} N(Y_i \cdot \mu, \Sigma)$, labels $Y_i = \pm 1$ with equal prob.

- $(p, n) = (5000, 1000)$
- $\sqrt{n}\mu(j) \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_\tau$, $(\epsilon, \tau) = (.1, 3.5)$
- Ω: tridiagonal (1 on diagonal, .4 on two sub-diagonals)
- 10 independent data splits

	HCT-PCS	HCT-glasso	SVM	RandForest
Average error	11.08	43.67	20.03	35.51
"Best" error	8.73	37.95	18.98	31.93

2.2. Phase diagrams for classification

Rare/Weak Feature Model

- n training samples $(X_i,\ Y_i)$: $X_i \sim N(Y_i \cdot \mu, \Sigma)$, $Y_i = \pm 1$ with equal prob.
- test sample: $X \sim N(\pm\mu,\ \Sigma)$
- summarizing Z-scores:

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i \cdot X_i) \sim N(\sqrt{n}\mu, \Sigma)$$

RW Feature Model, II

$$\sqrt{n} \cdot \mu(j) \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon \cdot \nu_\tau$$

- use p as the driving asymptotic parameter
- link (ϵ, τ) to p

$$\epsilon_p = p^{-\beta}, \qquad \tau_p = \sqrt{2r \log p}, \qquad 0 < \beta, r < 1$$

- link sample size n to p (3 types of growth):
  - (*No growth*): n is fixed
  - (*Slow growth*): $1 \ll n \ll p^\theta$, for any $\theta > 0$
  - (*Regular growth*): $n = p^\theta$ for some $\theta \in (0, 1)$

Sparse classification boundary ($\Omega = I_p$)

Jin (2009)

Define *standard phase function*

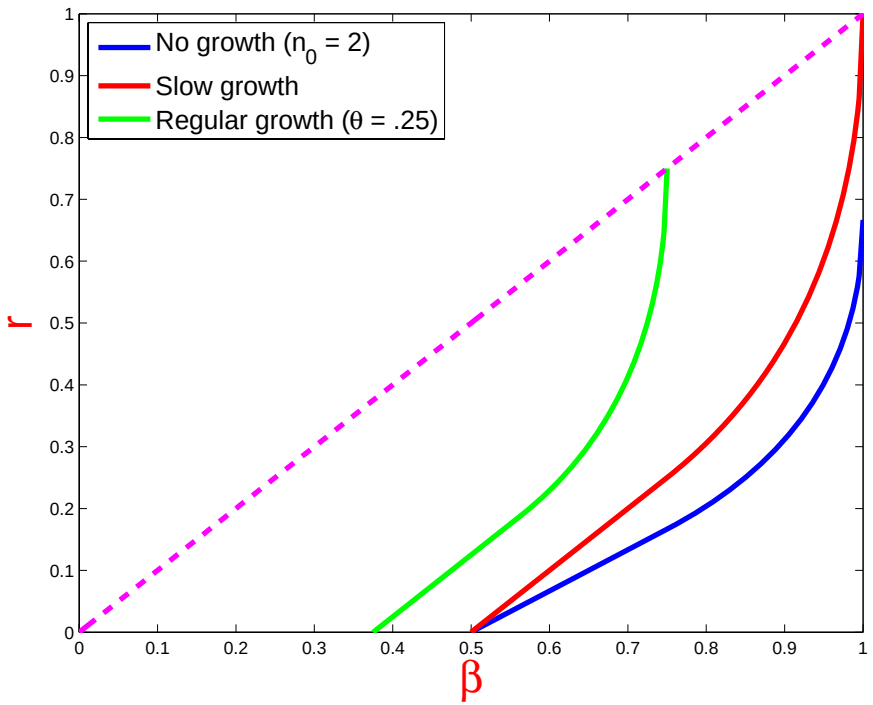
$$\rho(\beta) = \begin{cases} 0, & 0 < \beta < 1/2 \\ (\beta - 1/2), & 1/2 \le \beta < 3/4 \\ (1 - \sqrt{1-\beta})^2, & 3/4 \le \beta < 1 \end{cases}$$

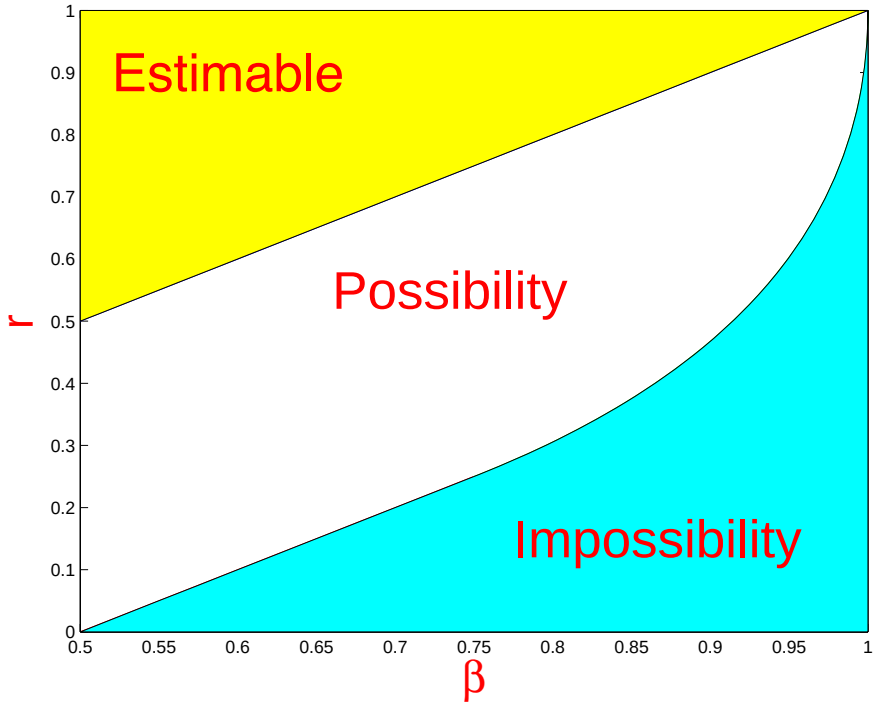
Let

$$(\star): \qquad r = \begin{cases} \frac{n}{n+1} \cdot \rho(\beta), & \text{no growth} \\ \rho(\beta), & \text{slow growth} \\ (1-\theta) \cdot \rho(\frac{\beta}{1-\theta}), & \text{regular growth} \end{cases}$$

Call $(\star)$ the *classification boundary*. It partitions the β-r plane into **Region of Possibility** and **Region of Impossibility**

Phase Diagram ($\Sigma = I_p$)





Left: Classification Boundaries. Right: Phase diagram (slow growth)
$\epsilon_p = p^{-\beta}$, $\tau_p = \sqrt{2r \log p}$, $0 < \beta, r < 1$

Summary of Part II

- Adapted Fisher's LDA to Rare/Weak settings
  - Estimating Ω from data by PCS (say)
  - Adapting the idea of HC for threshold choice in feature selection
- Clarified several popular ideas/beliefs
- Identified classification limits for RW model, and proved optimality of HC threshold choice

Main references

- Donoho D and Jin J (2008) Higher Criticism thresholding: optimal feature selection when useful features and rare and weak. *Proc. Natl. Acad. Sci.* **105**(39), 14790-14795.
- Jin J (2009) Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci.* **106**(22), 8859-8864.
- Donoho D and Jin J (2009) Feature selection by Higher Criticism thresholding: optimal phase diagram. *Phil. Trans. Roy. Soc. A.* **367**, 4449-4470.
- Fan Y, Jin J and Yao Z (2013) Optimal classification in sparse Gaussian graphic model. *Ann. Statist.* **41**(5), 2537-2571.
- Huang S, Jin J and Yao Z (2016) Partial Correlation Screening for estimating large precision matrices, with applications to classification. *Ann. Statist.* **44**(5), 2018-2057.

Part III: Clustering

Clustering

- Data (X_i, Y_i)
  - $Y_i \in \{1, 2, \cdots, K\}$: class labels
  - $X_i \sim N(\mu_{Y_i}, \Sigma)$: feature vectors in $\mathbb{R}^p$
- **Goal**. Given the sample matrix X, estimate the class label vector Y
- Motivation. Recover the underlying structure
  - Medicine: find the hidden groups of a disease, given the evaluations of the disease, and apply proper treatment
  - Biology: find possible subspecies of a species (iris flower data set example)

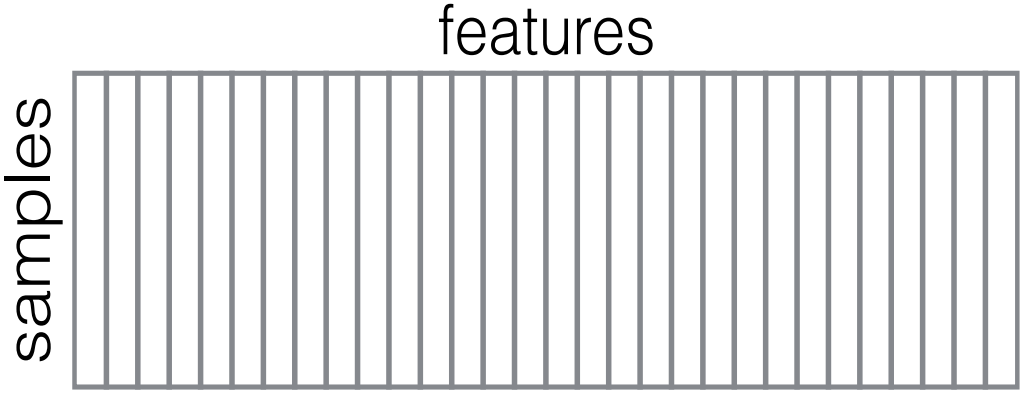
Clustering subjects using microarray data

#	Data Name	Source	K	n (# of subjects)	p (# of genes)
1	Brain	Pomeroy (02)	5	42	5597
2	Breast Cancer	Wang et al. (05)	2	276	22215
3	Colon Cancer	Alon et al. (99)	2	62	2000
4	Leukemia	Golub et al. (99)	2	72	3571
5	Lung Cancer	Gordon et al. (02)	2	181	12533
6	Lung Cancer(2)	Bhattacharjee et al. (01)	2	203	12600
7	Lymphoma	Alizadeh et al. (00)	3	62	4062
8	Prostate Cancer	Singh et al. (02)	2	136	6033
9	SRBCT	Kahn (01)	4	63	2308
10	Su-Cancer	Su et al (01)	2	174	7909

Goal. Predict the class labels (clustering)

We normalize in a way so that each feature has 0 *mean and unit variance*

Rare and Weak signals



= **"signal" + "noise**

- Each row has a "signal" and a "noise" component
  - Let $\mu_1, \mu_2, \ldots, \mu_K$ be the K group mean vectors
  - For row i, "signal" $= \mu_k$ if $i \in$ group k (rank: $K-1$)
  - "Noise": sample variation, measurement noise, etc.
- Rare/Weak signals:

$$\sum_{k=1}^{K} \mu_k^2(j) = \begin{cases} 0, & \text{for most (feature) } j \\ \text{relatively small}, & \text{for other (feature) } j \end{cases}$$

Related work

- Low-dimensional clustering
  - Hierarchical Clustering
  - Classical k-means [James MacQueen (1967)]

$$\arg\min_{\{g_1,g_2,\ldots,g_K\}}\left\{\sum_{k=1}^{K}\left(\sum_{i\in g_k}\|X_i-\mu_k\|^2\right)\right\},$$

    where $\{1,2,\ldots,n\}=g_1\cup g_2\cup\ldots\cup g_K$ is a partition and μ_k is the mean of $\{X_i, i\in g_k\}$
  - Spectral Clustering [Lee at al (2010)]
- High dimensional clustering with sparsity
  - Sparse PCA [Zou, Hastie, Tibshirani (2006)]
  - Sparse k-means [Witten and Tibshirani (2010)]

3.1. Influential Features PCA

Pearson's PCA

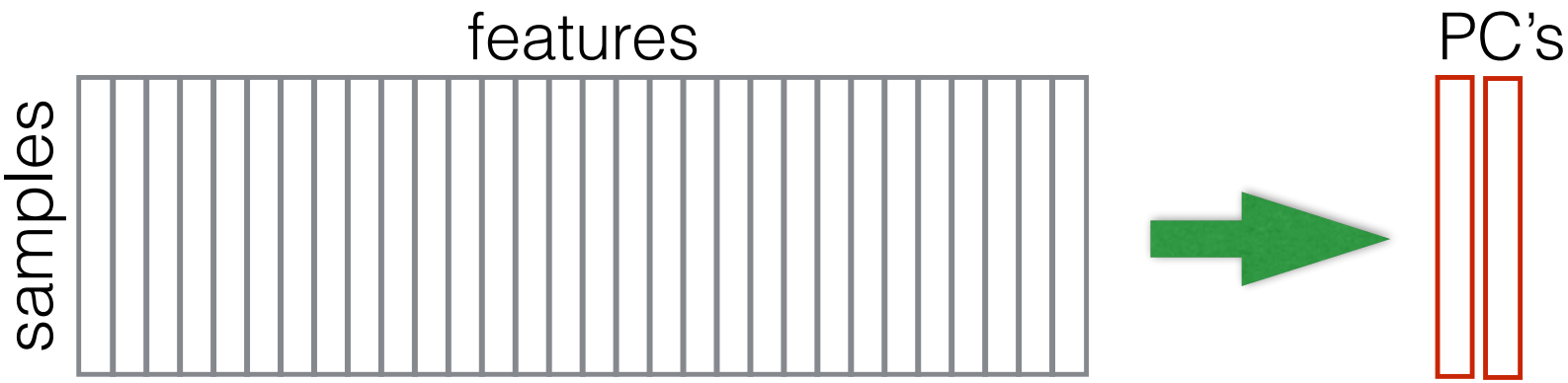
Karl Pearson (1857-1936)

- Transformation
- Dimension Reduction
- Noise Reduction

Cluster by classical PCA

Apply k-means to the matrix

$$[u_1, u_2, \ldots, u_{K-1}], \qquad u_k\text{: } k\text{-th left singular vector}$$



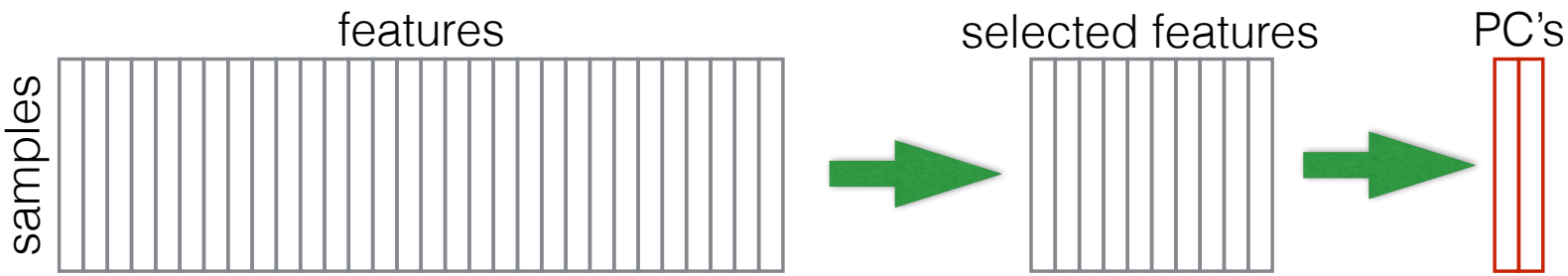
PC: Principal Components

- **Problem**. $\|$ "signal" $\| \ll \|$ "noise" $\|$ ($\|\cdot\|$: spectral norm)
- **Solution**. Feature selection

Influential Feature PCA (IF-PCA)

Idea: PCA applied to carefully selected features:

- Rank features by Kolmogorov-Smirnov statistic
- Retain features that are ranked on the top
- Apply PCA to the post-selection data matrix



Azizyan et al (2013), Chan and Hall (2010), Fan and Lv (2008)

IF-PCA (microarray data)

1. Rank features with KS-scores $\psi_n^{(1)}, \psi_n^{(2)}, \ldots, \psi_n^{(p)}$
2. Renormalize the KS-scores by (**Efron's empirical null**)

$$\psi_n^{*(j)} = \frac{\psi_n^{(j)} - \text{mean of all } p \text{ different } KS\text{-scores}}{\text{SD of all } p \text{ different } KS\text{-scores}}$$

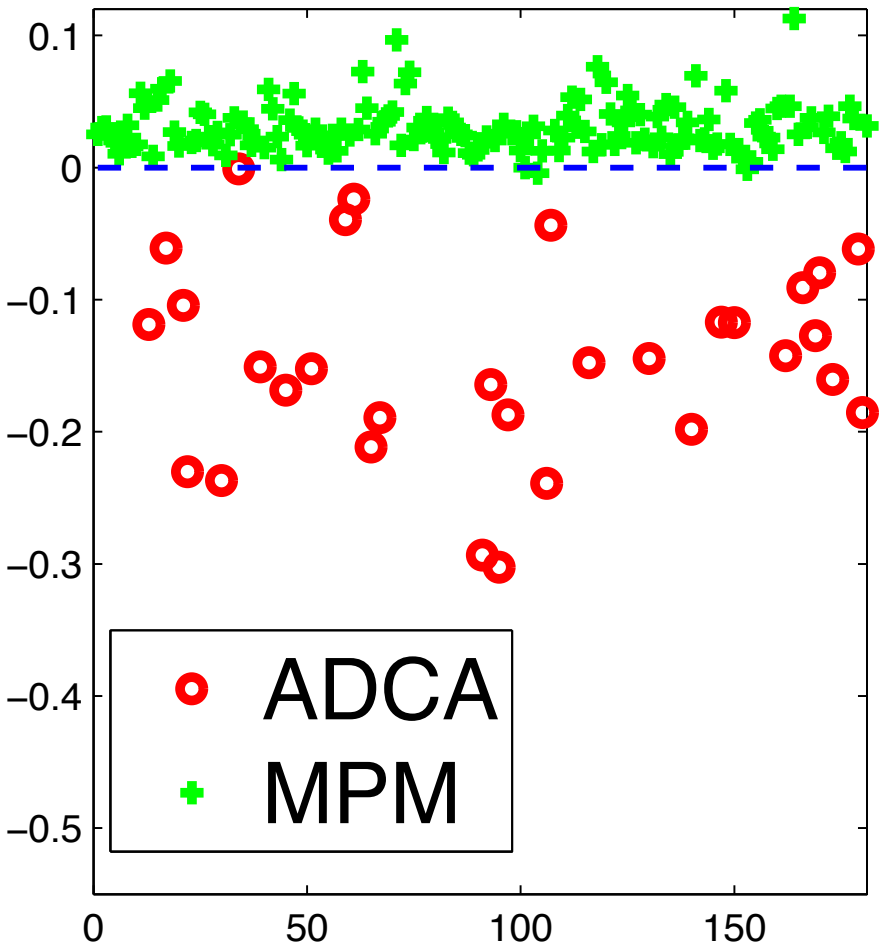
3. Apply PCA to post-selection mat. $\hat{X}^{(t)} = [x_j : \psi_n^{*(j)} \geq t]$
   - Let u_k be the k-th **left** singular vectors of $\hat{X}^{(t)}$
   - Apply k-means to $\hat{U}^{(t)} = [u_1, u_2, \ldots, u_{K-1}]$

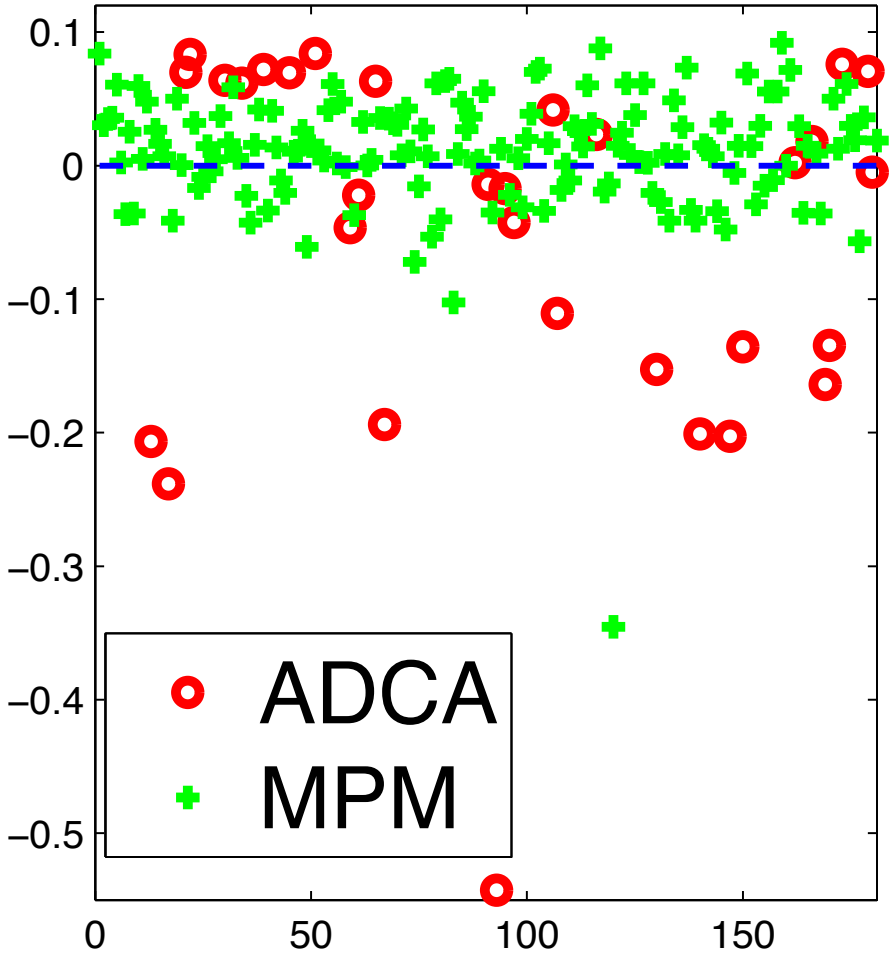
IF-PCA-HCT : $t = t_p^{HC}$: **H**igher **C**riticism threshold (**TBA**)

The blessing of feature selection

Lung Cancer (K = 2); x-axis: $1, \ldots, n$; y-axis: entries of $\hat{U}_{HC}$ $(t = t_p^{HC})$

Left: plot of $\hat{U}_{HC}$. Right: counterpart of $\hat{U}_{HC}$ without feature selection





Feature-wise Kolmogorov-Smirnov scores

$$X = [x_1, \ldots, x_p], \quad x_j \sim N(\eta_j, I_n)$$

$$\text{feature } j \text{ is useless} \iff \eta_j = 0$$

- Compute P-values: $\pi_i^{(j)} = P\big((N(0,1) \geq x_j(i)\big)$
- Sort P-values: $\pi_{(1)}^{(j)} < \pi_{(2)}^{(j)} < \ldots < \pi_{(n)}^{(j)}$
- Evaluate the significance of feature j by KS-score

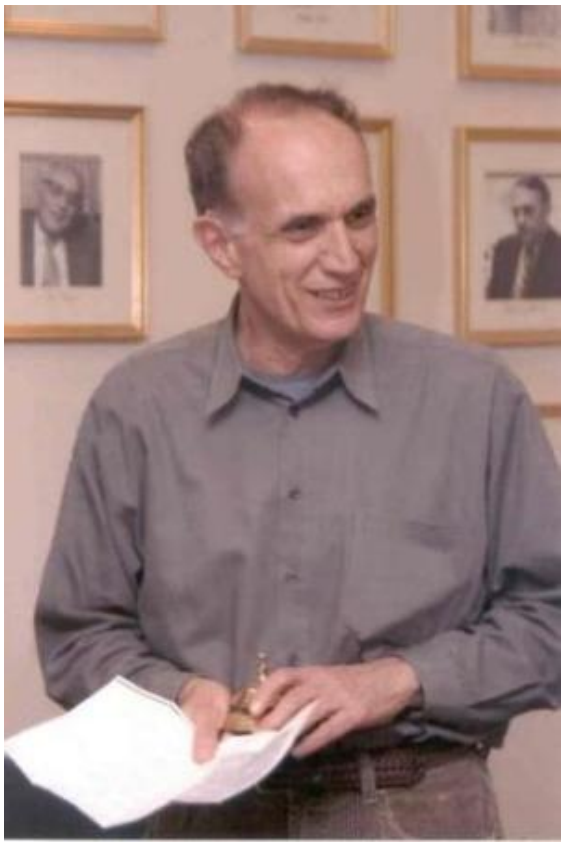
$$\psi_n^{(j)} = \max_{\{1 \leq k \leq n\}} \{|(k/n) - \pi_{(k)}^{(j)}|\}, \qquad j = 1, 2, \ldots, p$$

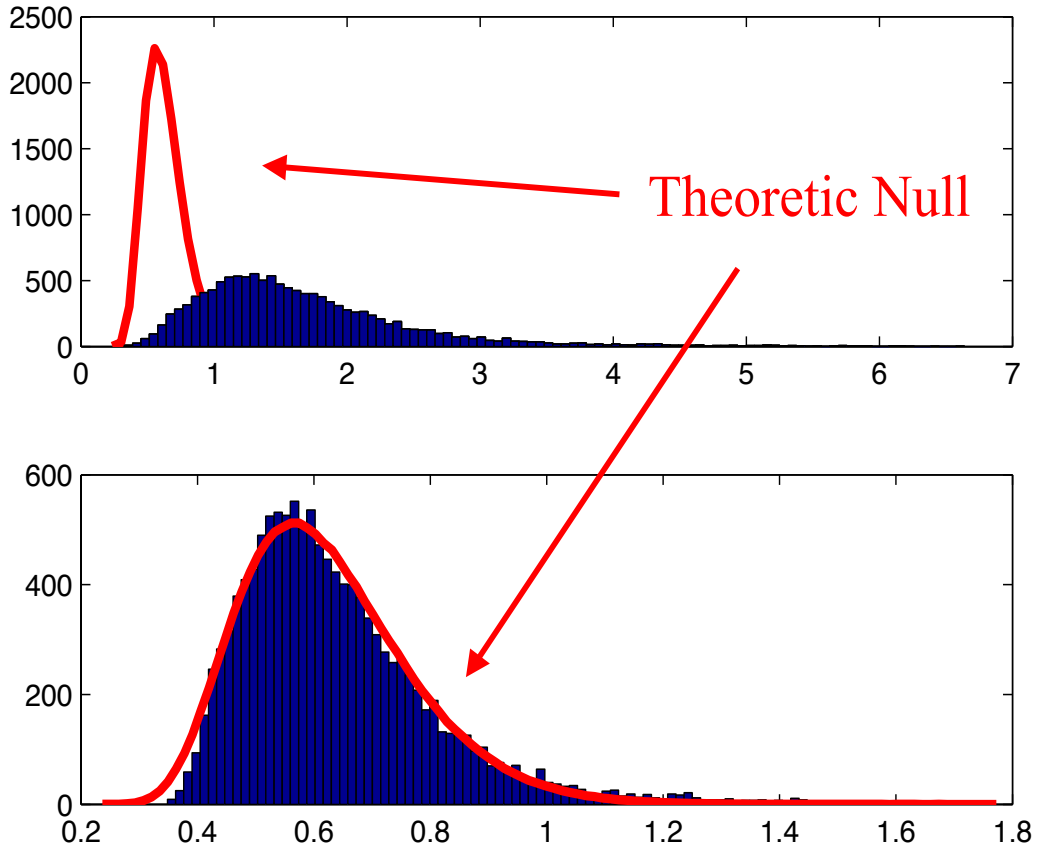
- Similar to HC, but signals here are non-sparse

Efron's null correction (Lung Cancer)

Theoretic Null: null density of KS-scores $\psi_n^{(j)}$ if data are Gaussian (does not depend on feature mean/variance, **easy to simulate**)

Theoretic null is a bad fit to $\psi_n^{(j)}$ (top) but a nice fit to $\psi_n^{*(j)}$ (bottom)





How to set the threshold t?

- CV: not implementable (class labels unknown)
- FDR: need tuning and target on **Rare/Strong** signals *[Benjamini and Hochberg (1995), Efron (2010)]*

t (threshold)	#{selected features}	feature-FDR	errors
.0280	12529	1.00	22
.1595	2523	1.00	28
.2814	**299**	**.538**	**4**
.2862	**280**	**.50**	**5**
.3331	**132**	**.25**	**6**
.3469	106	.20	43
.3622	86	.15	38
.4009	32	.10	38
.4207	27	.06	37

Threshold choice by HC

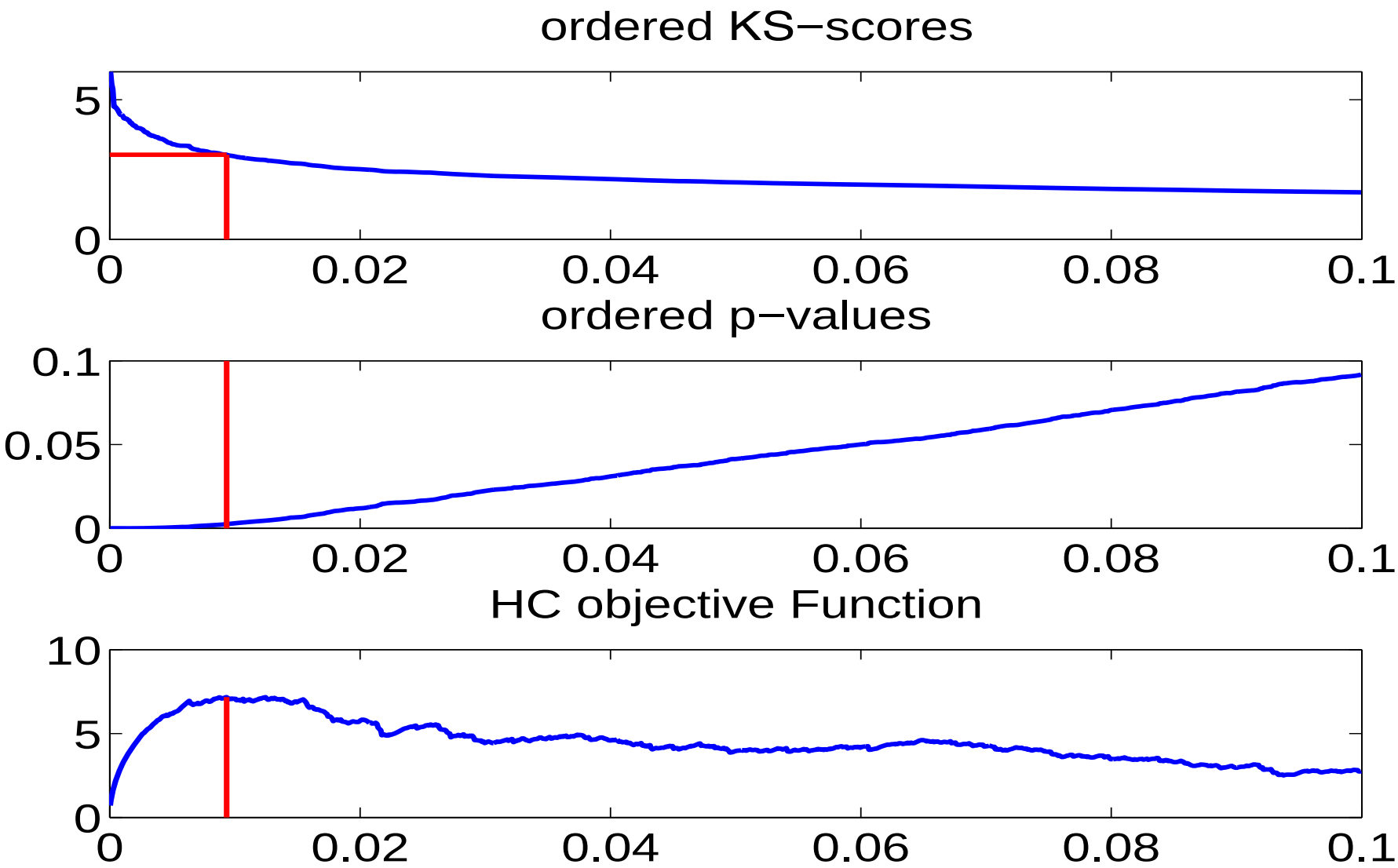
Jin and Wang (2015), Jin, Ke and Wang (2015)

- Compute P-values: $\pi_j = 1 - F_0(\psi_n^{*(j)})$, $1 \le j \le p$ (F_0: CDF of Efron's theoretical null)
- Sort P-values: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$
- Define

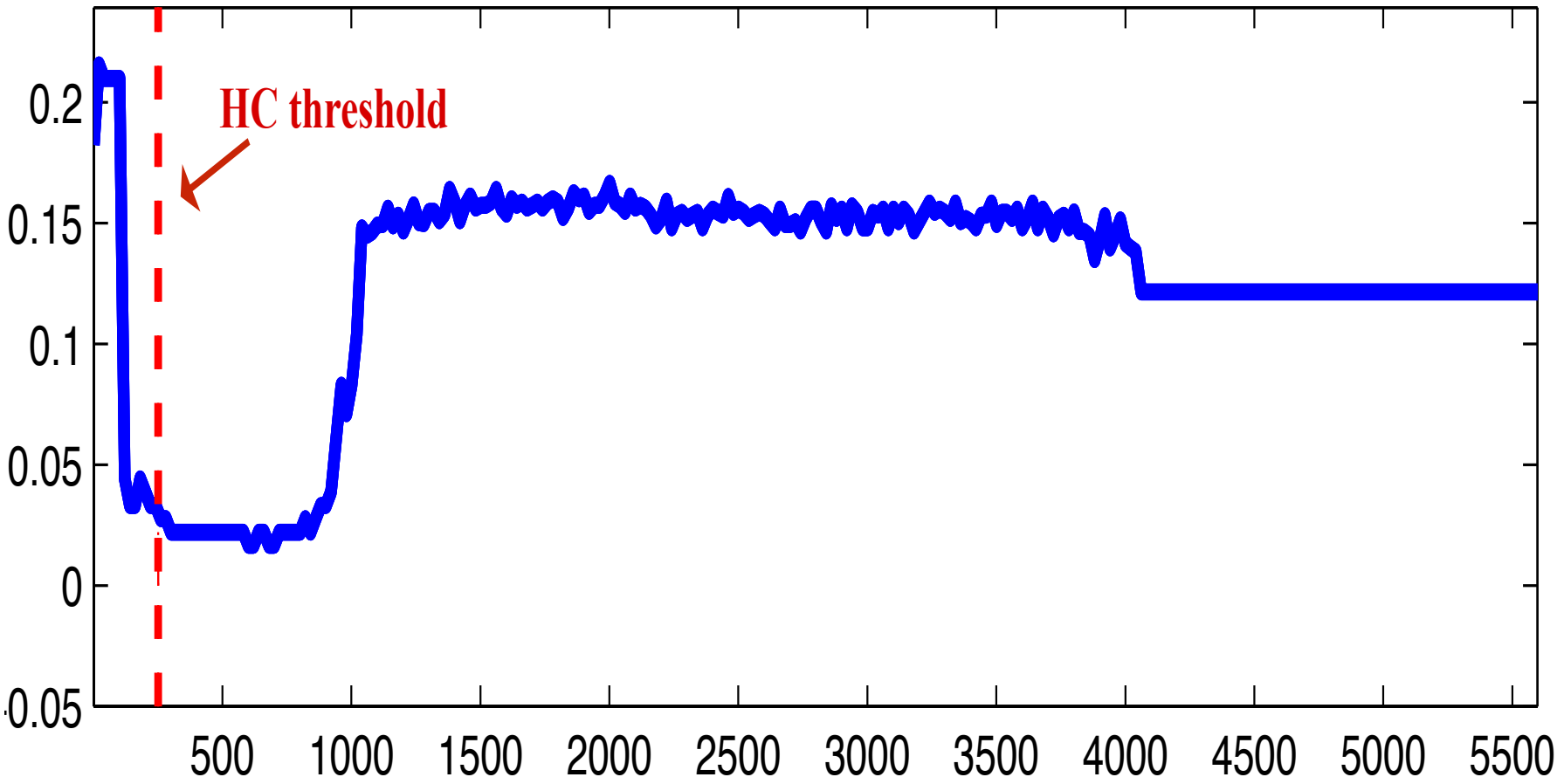
$$HC_{p,k} = \frac{\sqrt{p}\big(k/p - \pi_{(k)}\big)}{\sqrt{k/p + \max\{\sqrt{n}\big(k/p - \pi_{(k)}\big), 0\}}}$$

- Let $\hat{k} = \mathrm{argmax}_{\{1 \le k \le p/2, \pi_{(k)} > \log(p)/p\}} \{HC_{p,k}\}$. HC threshold t_p^{HC} is the $\hat{k}$-th largest KS-score

Illustration



Illustration, II (Lung Cancer)



x-axis: # of selected features; y-axis: error rate by IF-PCA

Comparison

- SpecGem: modified classical PCA [Lee et al (2010)]
- kmeans++: Authur and Vassilvitskii (2007)
- COSA: Friedman and Meulman (2004)
- skmeans: Witten and Tibshirani (2010)
- sPCA: Zou, Hastie, Tibshirani (2006) (project data matrix to the estimated feature space then cluster; tuning parameter ideally set)

#	Data set	K	kmean	SpecGem	kmean++	COSA	skmean	sPCA	Our
1	Brain	5	.286	.143	.427	.405	.286	.262	.262
2	Breast Cancer	2	.442	.438	.430	.359	.442	.438	.406
3	Colon Cancer	2	.443	.484	.460	.408	.468	.435	.403
4	Leukemia	2	.278	.292	.257	.167	.278	.292	.069
5	Lung Cancer	2	.116	.122	.196	.011	.116	.110	.033
6	Lung Cancer(2)	2	.436	.434	.439	.350	.448	.433	.217
7	Lymphoma	3	.387	.226	.433	.371	.387	.190	.065
8	Prostate	2	.422	.422	.432	.412	.422	.422	.382
9	SRBCT	4	.556	.508	.524	.587	.556	.428	.444
10	SuCancer	2	.477	.489	.459	.328	.477	.437	.333

Comparison, II

$$\text{Regret} = \frac{\text{cell value} - \text{row min}}{\text{row max} - \text{row min}}$$

#	Data set	K	kmean	SpecGem	kmean++	COSA	skmean	sPCA	Our
1	Brain	5	.504	0	1	.923	.504	.419	.419
2	Breast Cancer	2	1	.952	.855	0	1	.952	.566
3	Colon Cancer	2	.494	1	.704	.062	.803	.395	0
4	Leukemia	2	.937	1	.843	.440	.937	1	0
5	Lung Cancer	2	.568	.600	1	0	.568	.535	.119
6	Lung Cancer(2)	2	.948	.939	.961	.576	1	.945	0
7	Lymphoma	3	.875	.438	1	.832	.875	.340	0
8	Prostate	2	.800	.800	1	.600	.800	.800	0
9	SRBCT	4	.805	.503	.604	1	.805	0	.100
10	SuCancer	2	.926	1	.814	0	.926	.677	.031

	kmean	SpecGem	kmean++	COSA	skmean	sPCA	Our
Overall ranking	5	4	7	2	6	3	1
Average regret	.786	.723	.878	.443	.821	.605	.124
Average error rate	.384	.356	.406	.340	.388	.345	.261

3.2. Phase transitions in clustering

A Rare/Weak model ($K = 2$)

$$X = \ell\mu' + Z \quad \in \mathbb{R}^{n,p}, \qquad Z\text{: } \textit{iid } N(0,1) \text{ entries}$$

- Class labels:

$$\ell_i = \pm 1 \qquad \text{with equal prob.}$$

- Rows of "signal matrix" are $\pm\mu'$ (two classes)
- Feature strengths:

$$\mu(j) = \begin{cases} 0, & \text{with prob. } (1-\epsilon) \\ \tau, & \text{with prob. } \epsilon \end{cases}$$

Calibrations

We use p as driving asymptotic parameter and tie (n, ϵ, τ) to p by fixed parameters $(\theta, \beta, \underline{\alpha, r})$

- "large p small n":

$$n = p^{\theta}, \qquad 0 < \theta < 1$$

- Rare signals:

$$\#\{\text{of signals}\} \;\; s_p \approx p\epsilon_p = p^{1-\beta}, \qquad 0 < \beta < 1$$

- Weak signals:

$$\tau_p = p^{-\alpha} \cdot \sqrt[4]{4r \log(p)}$$

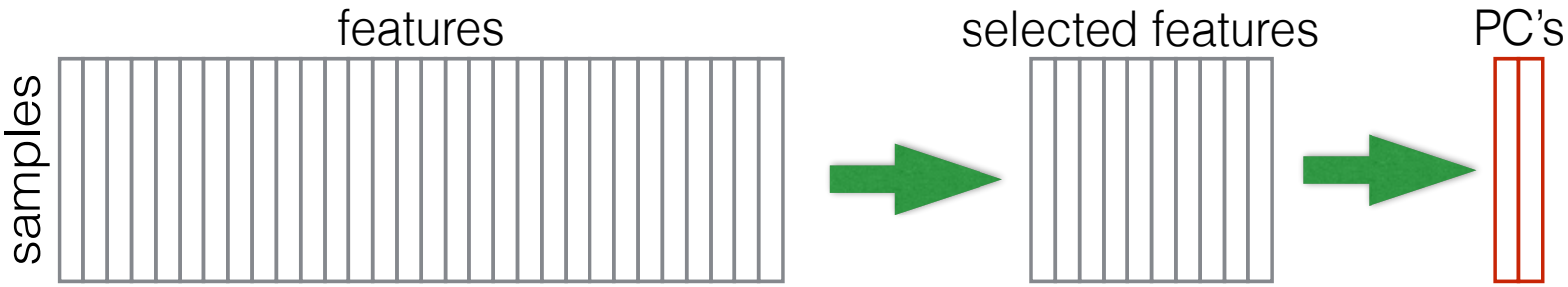
Two clustering limits

$$n = p^{\theta}, \qquad s_p \approx p\epsilon_p = p^{1-\beta}, \qquad \tau_p = p^{-\alpha}\sqrt[4]{4r\log(p)}$$

- **Statistical limits**. Fixing θ, for which (β, α) successful clustering is possible/impossible?
- **Computable statistics limits**. Fix θ and restrict to methods computable in polynomial time. For which (β, α), successful clustering is possible/impossible?

We study these in a **cruder** scale, so $\sqrt[4]{4r\log(p)}$ is negligible

IF-PCA for synthetic Gaussian data

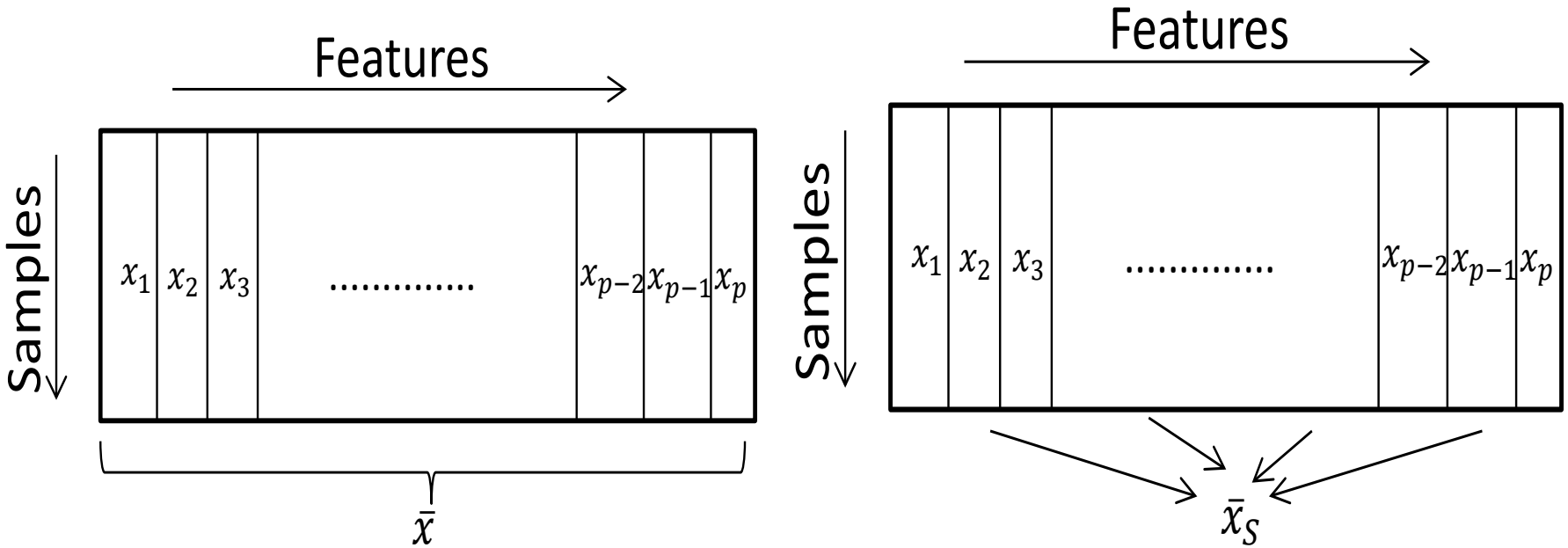


	microarray	two-class model
pre-normalization	yes	**skipped**
feature-wise screening	Kolmogorov-Smirnov scores $\psi_n^{(j)}$	**chi-square** scores $\psi_n^{(j)} = (\|x_j\|^2 - n)/\sqrt{2n}$
re-normalization	Efron's null correction	**skipped**
threshold choice	HCT	same
post-selection PCA	same	same

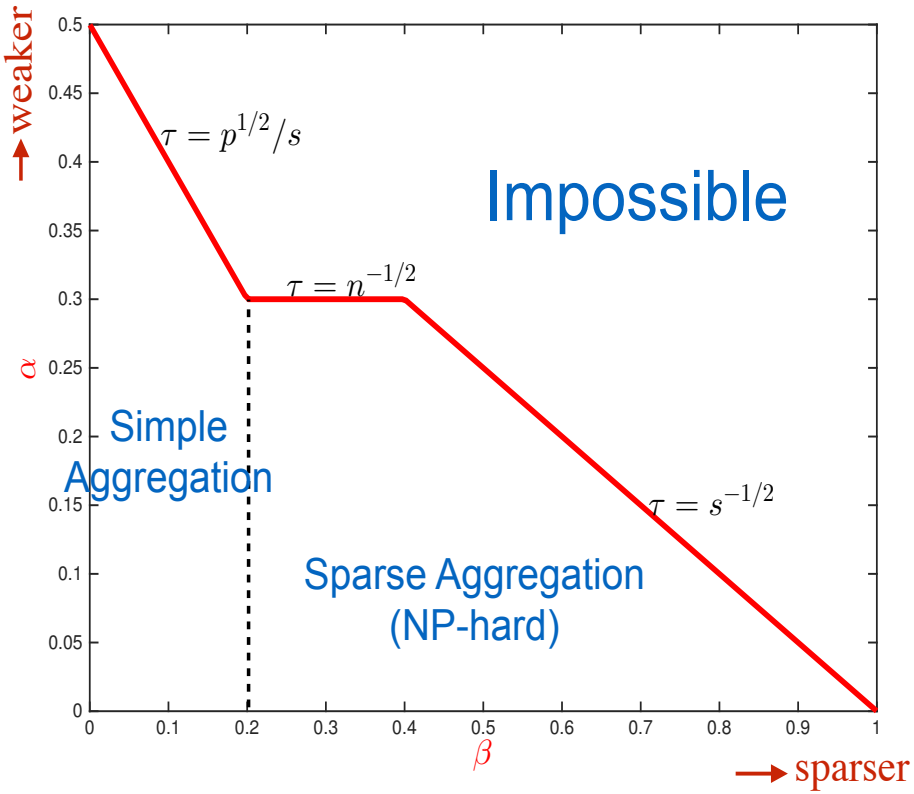
Two aggregation methods

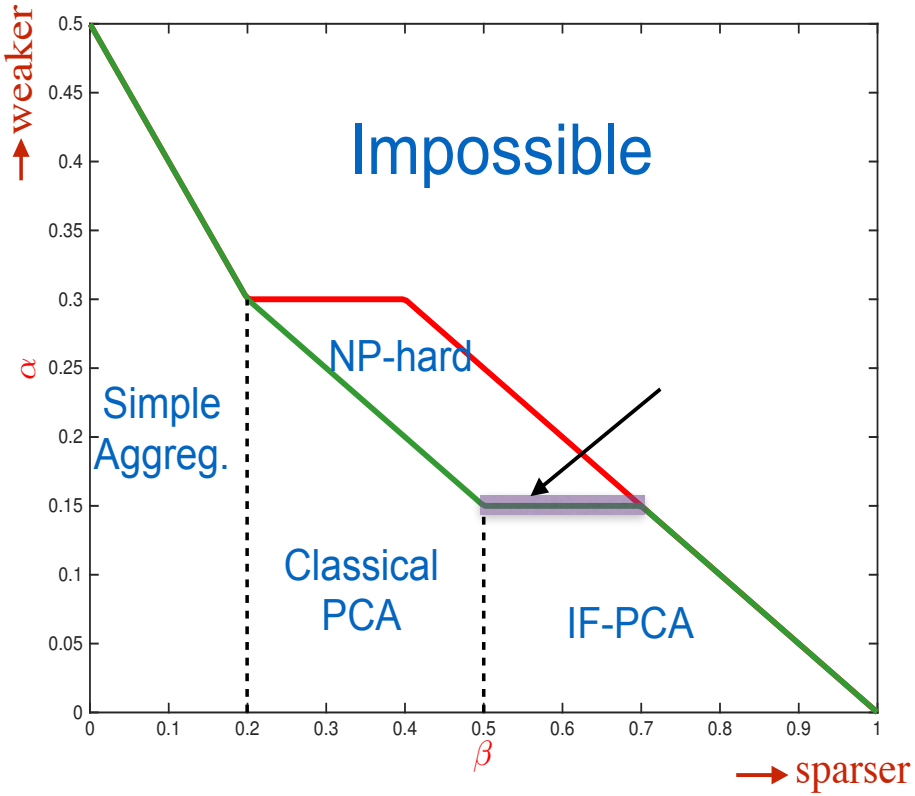
- Simple Aggregation: cluster by $\operatorname{sgn}(\bar{x})$
- Sparse Aggregation: for a tuning integer N, clusters by $\operatorname{sgn}(\bar{x}_{\hat{S}})$, where

$$\hat{S} = \operatorname{argmax}_{\{S:|S|=N\}} \left\{ \|\bar{x}_S\|_1 \right\}$$



Two limits (β-α plane, cruder scale)





Sparsity level	Signals	Optimal methods	Feature Selection
$\sqrt{np} \ll s_p \ll p$	Very weak	Simple Agg.	Impossible
$\sqrt{p} \ll s_p \ll \sqrt{np}$	Very weak	PCA	Impossible
$\sqrt{n} \ll s_p \ll \sqrt{p}$	Critical	IF-PCA	Necessary & subtle
$1 \ll s_p \ll \sqrt{n}$	Strong	IF-PCA	Necessary yet trivial

IF-PCA: most interesting on "shaded bar"

$$n = p^\theta, \qquad 1/2 < \beta < 1 - \theta/2, \qquad \tau_p = p^{-\alpha}\sqrt[4]{4r\log(p)}$$

Fixing θ, $\alpha = \frac{\theta}{4}$ is a constant, $\sqrt[4]{4r\log(p)}$ plays a key role:

$$\chi^2\text{-score } \frac{\|x_j\|^2 - n}{\sqrt{2n}} \approx \begin{cases} N(0,\ 1), & \text{feature } j \text{ useless} \\ N(\sqrt{2r\log(p)},\ 1), & \text{feature } j \text{ useful} \end{cases}$$

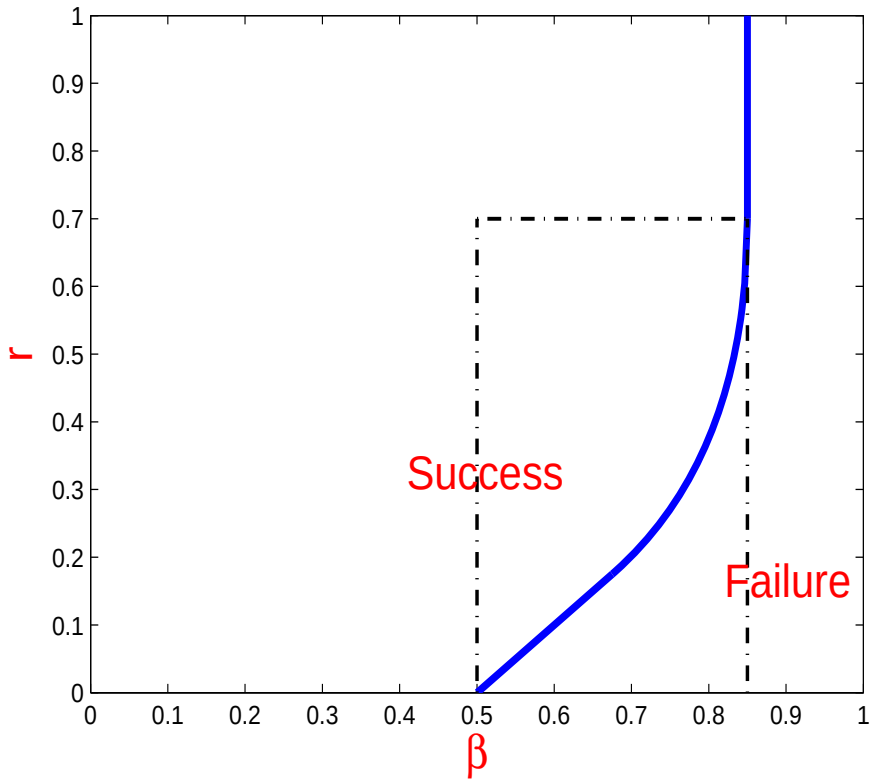
- $r \ll 1$: feature selection impossible (signals too weak)
- $r > 1$: feature selection is easy (signals too strong)
- **Limits for IF-PCA**. Fix θ. For which (β, r), successful clustering by IF-PCA is possible/impossible?

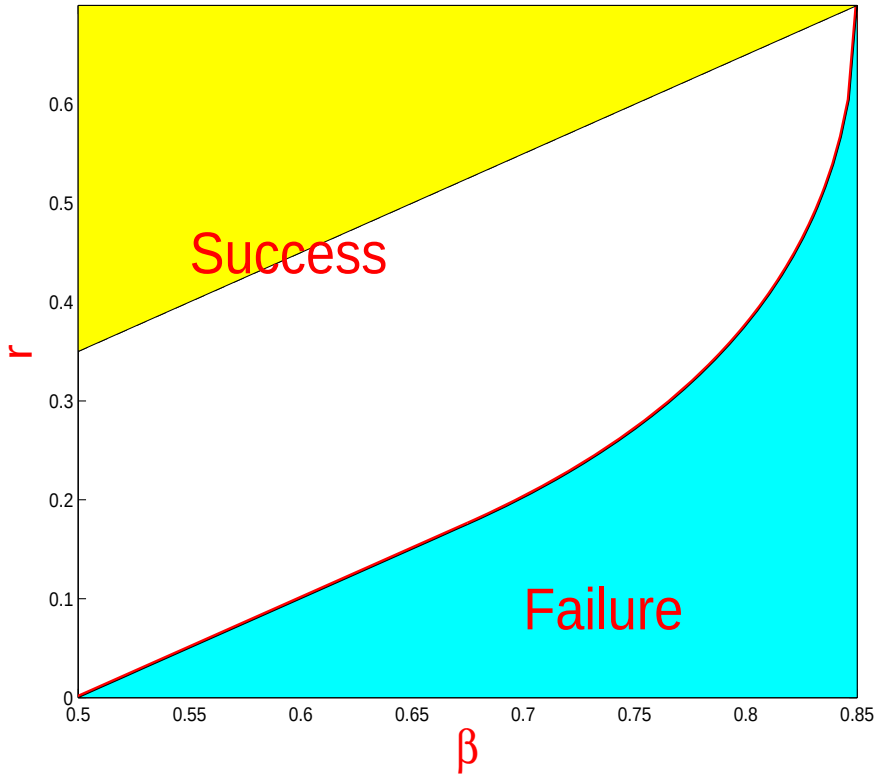
Phase transition for IF-PCA (β-r plane)

Recall that $n = p^\theta$. Define the standard phase function:

$$\rho(\beta) = \begin{cases} \beta - \frac{1}{2}, & \frac{1}{2} < \beta < \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2, & \frac{3}{4} < \beta < 1 \end{cases}$$

and let $\rho_\theta(\beta) = (1-\theta)\rho(\frac{1}{2} + \frac{\beta - \frac{1}{2}}{1-\theta}), \frac{1}{2} < \beta < 1 - \frac{\theta}{2}$





Consistency of HC threshold

Consider IF-PCA with a threshold $t > 0$. Let $Err(t)$ be the clustering error, and $\hat{\xi}^{(t)}$ be the first singular vector of the post-selection data matrix. Fix (β, r, θ) such that $r > \rho_\theta(\beta)$.

- There is a non-stochastic function $\widetilde{SNR}(t)$ such that for t in an appropriate range, $\widetilde{SNR}(t) \gg 1$ and

$$\hat{\xi}^{(t)} \propto \widetilde{SNR}(t)\ell + z + rem, \qquad z \sim N(0, I_n)$$

- Define two non-stochastic thresholds:

$$t_p^{idealHC} = \mathrm{argmax}_t \widetilde{SNR}(t), \qquad t_p^{ideal} = \mathrm{argmax}_t Err(t).$$

  - HCT yields the right threshold choice:

$$t_p^{HC}/t_p^{idealHC} \to 1 \text{ in prob.}, \qquad t_p^{idealHC}/t_p^{ideal} \to 1$$

  - IF-PCA-HCT yields successful clustering

Summary of Part III

- Proposed IF-PCA for cancer subject clustering
  - Feature ranking by Kolmogorov-Smirnov statistics
  - Reduce gaps between theory and data analysis by Efron's null correction
  - Threshold choice by Higher Criticism
- IF-PCA is conceptually simple, flexible, tuning free, and an example of "Screen first, Estimate next" philosophy
- Identified statistical limits and computationally tractable statistical limits for clustering problem
- Proved consistency of Higher Criticism threshold and that IF-PCA-HCT achieves the optimal phase diagram

References

- Jin J, Wang W (2015) Influential Features PCA for high dimensional clustering (with discussion). *Ann. Statist.* **44**(6), 2323-2359.
- Jin J, Ke Z, Wang W (2015) Phase transitions for high dimensional clustering and related problems. *Ann. Statist*, **45**(5), 2151-2189.

Part IV: Variable Selection

Sparse linear regression model

$$Y = X\beta + z, \qquad X \in \mathbb{R}^{n,p}, \qquad z \sim N(0, I_n)$$

- Both p and n are large
- Signals are Rare/Weak
- Columns of X are normalized, and each column is only significantly correlated with a few others

Variable selection approaches

- Stage-wise methods (Forward, Forward-Backward)
- Penalization methods
  - Subset selection:

$$\frac{1}{2}\|Y - X\beta\|^2 + \frac{1}{2}\lambda^{ss}\|\beta\|_0$$

  - The lasso:

$$\frac{1}{2}\|Y - X\beta\|^2 + \lambda^{lasso}\|\beta\|_1$$

- Screen-and-Clean methods (**Today**):
  - Screen: remove a large fraction of noise variables by a computationally efficient algorithm
  - Clean: further filter out falsely selected variables

4.1. Non-optimality of penalization methods

Measure of success

- Oracle property (inappropriate for RW signals)†

$$P\{S(\hat{\beta}) = S(\beta)\}, \qquad (S(\beta) = \{1 \le j \le p : \beta_j \neq 0\}\text{: support})$$

- ℓ^q-norm (more relevant for prediction):

$$\|\hat{\beta} - \beta\|_q, \qquad q > 0$$

- Hamming selection error:

$$\mathrm{Hamm}_p(\hat{\beta}, \beta; \Omega) = E\left[\sum_{j=1}^{p} 1\{\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)\}\right]$$

† *Fan and Lv (2008)*

A Rare/Weak model and three regions

$$\beta_j = \begin{cases} 0, & 1-\epsilon_p \\ \tau_p, & \epsilon_p \end{cases},$$

$$\epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log(p)}, \qquad 0 < \vartheta, r < 1$$

- **Region of No recovery**: fixing (ϑ, r) in this region,

$$\mathrm{Hamm}_p(\hat{\beta}, \vartheta, r; X) \gtrsim p\epsilon_p$$

- **Region of Almost Full Recovery**

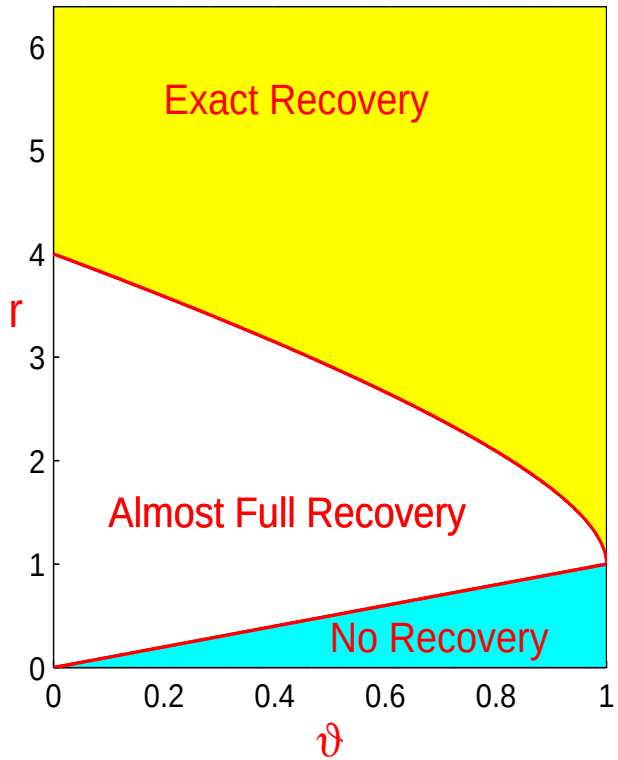
$$1 \ll \mathrm{Hamm}_p(\hat{\beta}, \vartheta, r; X) \ll p\epsilon_p$$

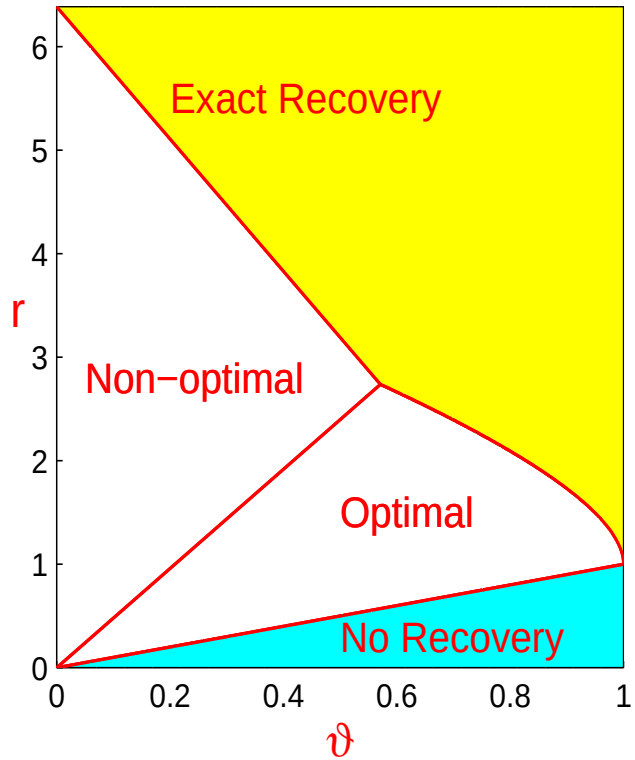
- **Region of Exact recovery**.

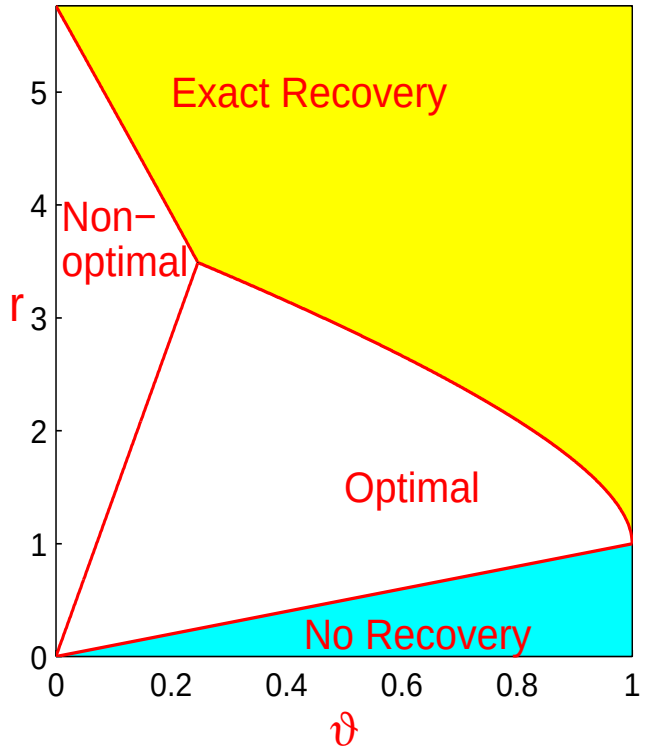
$$\mathrm{Hamm}_p(\hat{\beta}, \vartheta, r; X) = 0, \quad \text{except a small probability}$$

Phase Transition ($X'X$ nearly tridiag)

$$Y \sim N(X\beta, I_n), \quad \text{rows of } X \textit{ iid} \text{ from } N(0, \tfrac{1}{n}\Omega), \quad \Omega = \begin{bmatrix} \mathbf{1} & \mathbf{a} & 0 & \dots & 0 & 0 \\ \mathbf{a} & \mathbf{1} & \mathbf{a} & \dots & 0 & 0 \\ 0 & \mathbf{a} & \mathbf{1} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{a} & \mathbf{1} \end{bmatrix}$$







4.2. Covariate-assisted screening

Ranking variables

- Goal: rank the variables such that signal variables are ranked above most non-signals
- Useful for identifying a few most promising variables for follow-up lab experiments
- Useful for the follow-up tasks such as variable selection or FDR controlling

Marginal ranking

- Rank variables by

$$T_j = \left|\frac{(x_j, Y)}{(x_j, x_j)}\right|^2 = (x_j, Y)^2, \quad x_j\text{: } j\text{-th column of } X$$

- $P_{\mathcal{I}}$: projection from $\mathbb{R}^n$ to $\mathrm{span}\{x_j, j \in \mathcal{I}\}$

$$T_j = \|P_{\mathcal{I}} Y\|^2 - \|P_{\mathcal{I}\setminus\{j\}} Y\|^2, \qquad (\mathcal{I} = \{j\})$$

Problem. Signal Cancellation

$$(x_j, Y) = \sum_{k=1}^{p} (x_j, x_k)\beta_j + (x_j, z) = \beta_j + \sum_{k:k\neq j, \beta_k \neq 0} (x_j, x_k)\beta_k + (x_j, z)$$

Exhaustive multivariate ranking

$$T_j^* = \max_{\{\mathcal{I}: j \in \mathcal{I}, |\mathcal{I}| \leq m\}} T_{j|\mathcal{I}}, \qquad T_{j|\mathcal{I}} = \|P_{\mathcal{I}} Y\|^2 - \|P_{\mathcal{I}\setminus\{j\}} Y\|^2$$

- Computationally infeasible:

$$\sum_{k=1}^{m} \binom{p}{k}$$

- Inefficiency: T_j^* is the max over $\sum_{k=1}^{m} \binom{p-1}{k-1}$ χ^2-tests, and is large even if β_j is a noise

  To rank above average noise terms, signals need to be stronger far more than necessary

Graph Of Strong Dependence (GOSD)

Define GOSD $\mathcal{G} = (V, E)$:

- $V = \{1, 2, \ldots, p\}$: each variable is a node
- Nodes i and j have an edge iff

$$\left|(x_i, x_j)\right| \geq \delta, \qquad (\delta = \frac{1}{\log(p)}, \text{ say})$$

- Under our assumptions, $\mathcal{G}$ is sparse

Covariate-Assisted Ranking (CAR)

Rank variables by

$$T_j^* = \max_{\mathcal{I} \in \mathcal{A}_j(m)} T_{j|\mathcal{I}}, \quad T_{j|\mathcal{I}} = \|P_{\mathcal{I}} Y\|^2 - \|P_{\mathcal{I}\setminus\{j\}} Y\|^2$$

$\mathcal{A}_j(m)$: size $\leq m$ connected subgraphs containing j

Let d be the maximum degree of $\mathcal{G}$.

$$\left| \cup_{j=1}^p \mathcal{A}_j(m) \right| \leq Cp(2.718d)^m \ll \sum_{k=1}^m \binom{p}{k}$$

A real example

Data: gene expression of human immortalized B cells ($(p, n) = (4238, 148)$; Nayak et al. (2009))

- Remove the first singular vector:

$$\text{Data} = \sum_{k=1}^{n} \sigma_k u_k v_k' = \sigma_1 u_1 v_1' + \underbrace{\sum_{k=2}^{n} \sigma_k u_k v_k'}_{\text{design matrix } X}$$

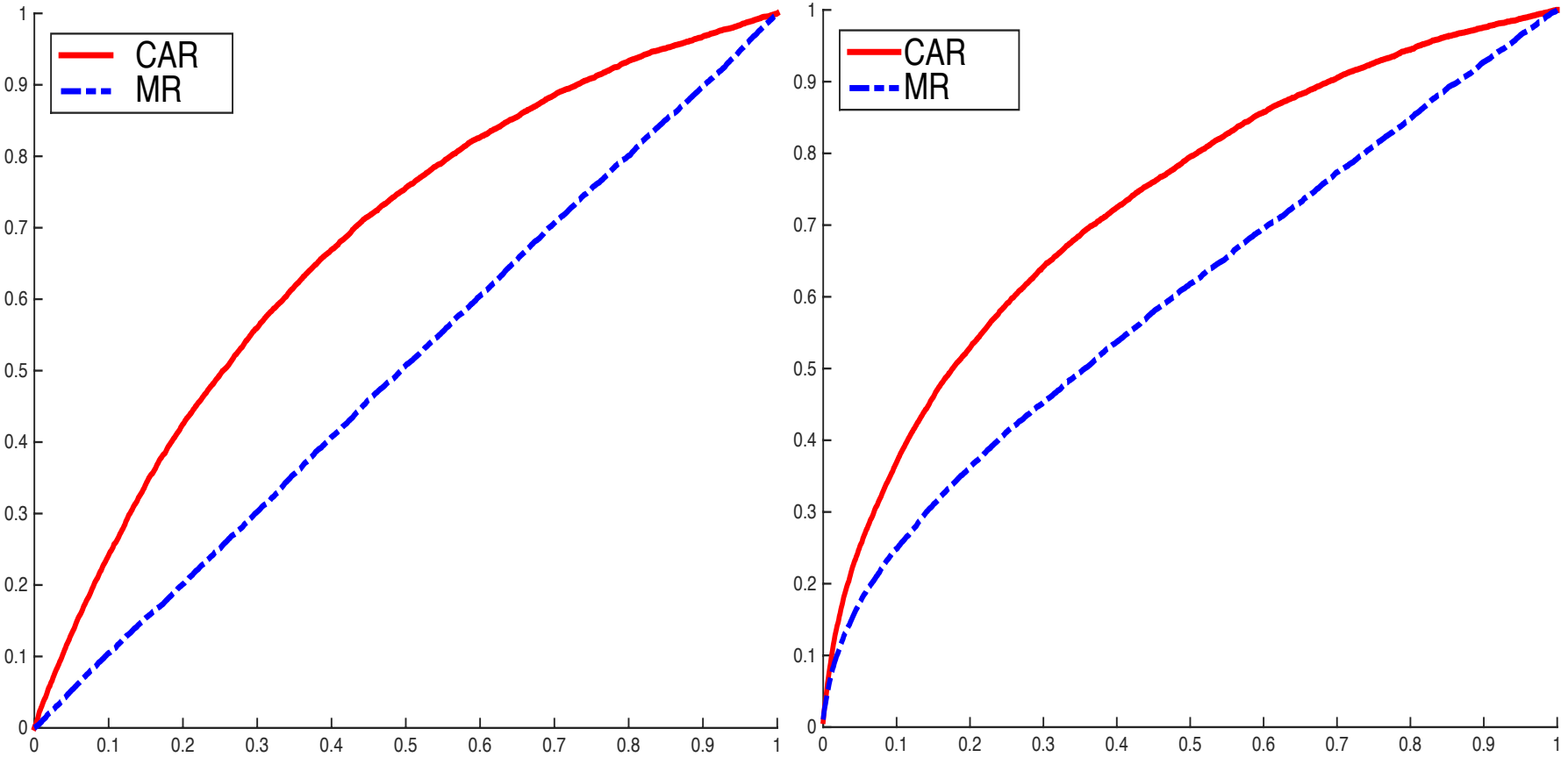
- Synthetic data for regression:

$$Y = N(X\beta, I_n), \qquad \beta_j \begin{cases} \sim N(0, \eta^2), & 1 \le j \le s \\ = 0, & \text{otherwise} \end{cases}$$

Comparison of the ROC curve

For CAR, $(m, \delta) = (2, 0.5)$ $\qquad \beta_j \begin{cases} \sim N(0, \eta^2), & 1 \le j \le s \\ = 0, & \text{otherwise} \end{cases}$

Left: $(\eta, s) = (0.1, 50)$. Right: $(\eta, s) = (5, 50)$



GOSD and mutual orthogonality

- Ranking is "easy" if $x_1, x_2, ..., x_p$ split into many **small-size mutually orthogonal groups**

$$\{1, 2, \ldots, p\} = \mathcal{I}_1 \cup \mathcal{I}_2 \cup \ldots \cup \mathcal{I}_m,$$

$$X^{\otimes, \mathcal{I}_k} \quad \perp \quad X^{\otimes, \mathcal{I}_\ell}, \qquad k \neq \ell$$

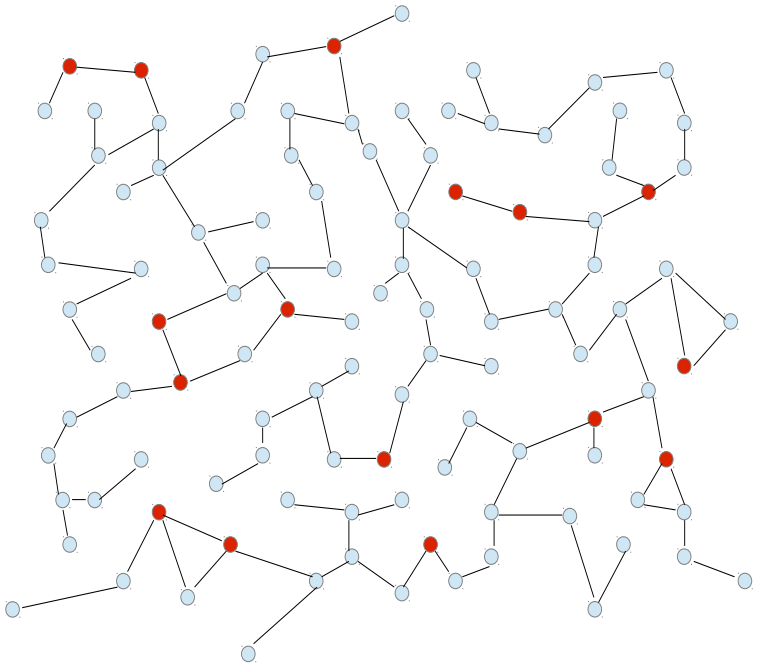
$X^{\otimes, \mathcal{I}}$: sub-matrix with columns restricted to $\mathcal{I}$
$\beta^{\mathcal{I}}$: sub-vector with rows restricted to $\mathcal{I}$

$$\underbrace{\mathcal{G} = \mathcal{G}_1 \cup \ldots \cup \mathcal{G}_M}_{\text{components}} \implies X^{\otimes, \mathcal{G}_k} \text{ approx. mutual. orth.}$$

Signal archipelago

$$\underbrace{\mathcal{G}_S}_{\text{signal subgraph}} = \underbrace{\mathcal{G}_{S,1} \cup \mathcal{G}_{S,2} \cup \ldots \cup \mathcal{G}_{S,M}}_{\text{components}}, \qquad S = S(\beta)$$

Lemma. If $I\{\beta_j \neq 0\} \overset{iid}{\sim} \text{Bernoulli}(\epsilon)$ and d is the max. degree of $\mathcal{G}$, then $P\left\{ \max_{1 \leq k \leq M} |\mathcal{G}_{m,k}| > m \right\} \leq p(ed\epsilon)^{m+1}$



Why signal cancellation is not an issue

$$\mathcal{G}_S = \mathcal{G}_{S,1} \cup \ldots \cup \mathcal{G}_{S,M}, \qquad T_{j|\mathcal{G}_{S,k_0}} = \|P_{\mathcal{G}_{S,k_0}} Y\|^2 - \|P_{\mathcal{G}_{S,k_0}\setminus\{j\}} Y\|^2$$

- Mutual orthogonality $\Rightarrow$ Signal cancellation only happens within $\mathcal{G}_{S,k_0}$
- For significance of β_j, it suffices to consider

$$Y = X^{\otimes,\mathcal{G}_{S,k_0}} \beta^{\mathcal{G}_{S,k_0}} + z \qquad (X^{\otimes,\mathcal{G}_{S,k_0}} \perp X^{\otimes,\mathcal{G}_{S,k}}, \text{approx.})$$

- $T_{j|\mathcal{G}_{S,k_0}}$ is optimal for testing

$$H_0 : \beta_j = 0 \qquad v.s. \qquad H_1 : \beta_j \neq 0$$

How CAR ensures successful ranking

$$T_j^* = \max_{\substack{\mathcal{I}:\ |\mathcal{I}| \le m\ \&\ \mathcal{I} \ni j \\ \text{connected subgraph}}} T_{j|\mathcal{I}}, \qquad \mathcal{G}_S = \mathcal{G}_{S,1} \cup \ldots \cup \mathcal{G}_{S,M}$$

- If j is a signal (say, $j \in \mathcal{G}_{S,k_0}$) and $m \ge |\mathcal{G}_{S,k_0}|$,

$$T_j^* \ge T_{j|\mathcal{G}_{S,k_0}} = \|P_{\mathcal{G}_{S,k_0}} Y\|^2 - \|P_{\mathcal{G}_{S,k_0}\setminus\{j\}} Y\|^2 \sim \chi_1^2(c^2\beta_j^2)$$

- If j is a noise,

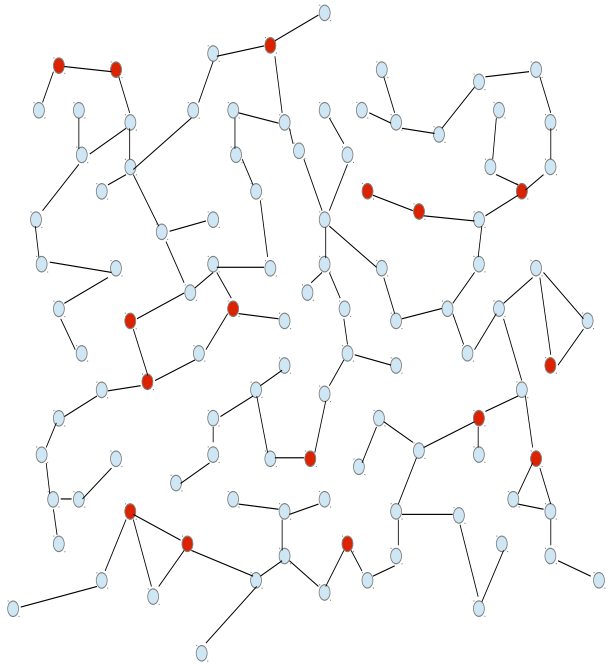
$$T_j^* = \begin{cases} \text{small}, & \text{if no signal in all } \mathcal{I} \\ \text{maybe large}, & \text{otherwise (Type I; not too many)} \end{cases}$$

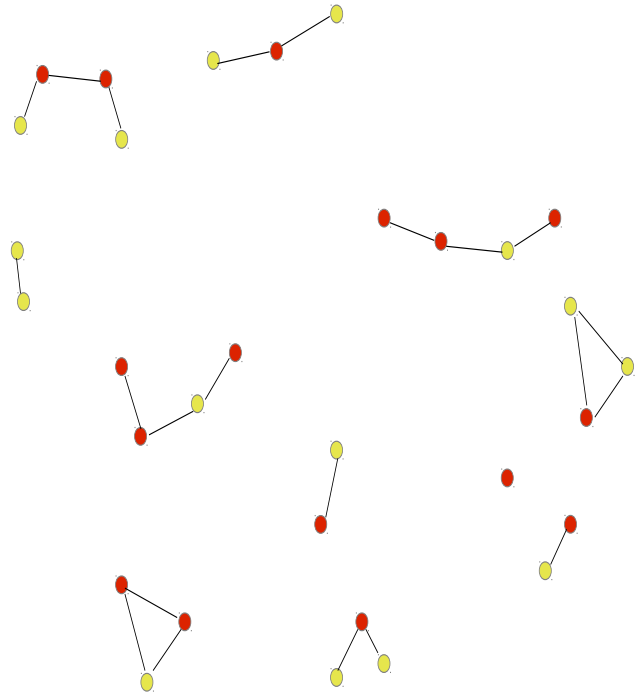
Therefore, signals are ranked above most non-signals!

4.3. Covariate Assisted Screening Estimation (CASE)

Post-screening archipelago

- CASE removes most noise while retain most signals
- Post-selection variables split into many **small-size** islands, each contains some signals and some noise
- **Keys**. sparsity of β, sparsity of $\mathcal{G}$, successful ranking





Variable selection by CASE

Fix tuning parameters (q, u^{case}, v^{case}), $t_q = \sqrt{2q\log(p)}$

- **Screen**. Rank variables by CAR and let

$$\hat{S}_t = \{1 \le j \le p,\ T_j^* \ge t_q\}$$

$$\underbrace{\mathcal{G}_{\hat{S}}}_{\text{post-selection subgraph}} = \underbrace{\mathcal{G}_{\hat{S},1} \cup \mathcal{G}_{\hat{S},2} \cup \ldots \cup \mathcal{G}_{\hat{S},\hat{M}}}_{\text{small-size components}}$$

- **Clean**.
  - If $j \notin \hat{S}_t$, set $\hat{\beta}_j = 0$.
  - Otherwise, we must have $j \in \mathcal{G}_{\hat{S},k}$ for some k. Estimate $\{\beta_j : j \in \mathcal{G}_{\hat{S},k}\}$ by minimizing

$$\|P_{\mathcal{G}_{\hat{S},k}}(Y - \sum_{j \in \mathcal{G}_{\hat{S},k}} \beta_j x_j)\|^2 + (u^{case})^2\|\beta\|_0, \qquad \underbrace{\beta_j = 0 \text{ or } \beta_j \ge v^{case}}_{\text{up to}}$$

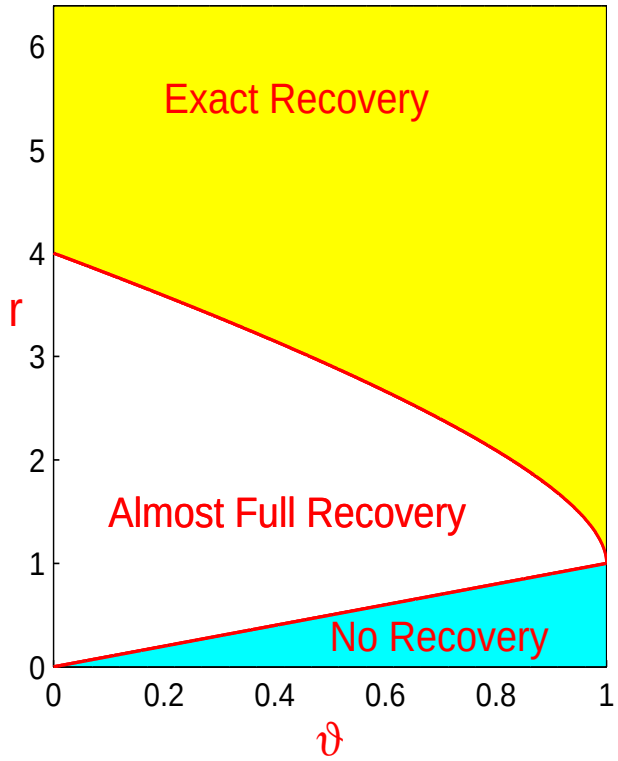
Comparison with L^0/L^1-penalization

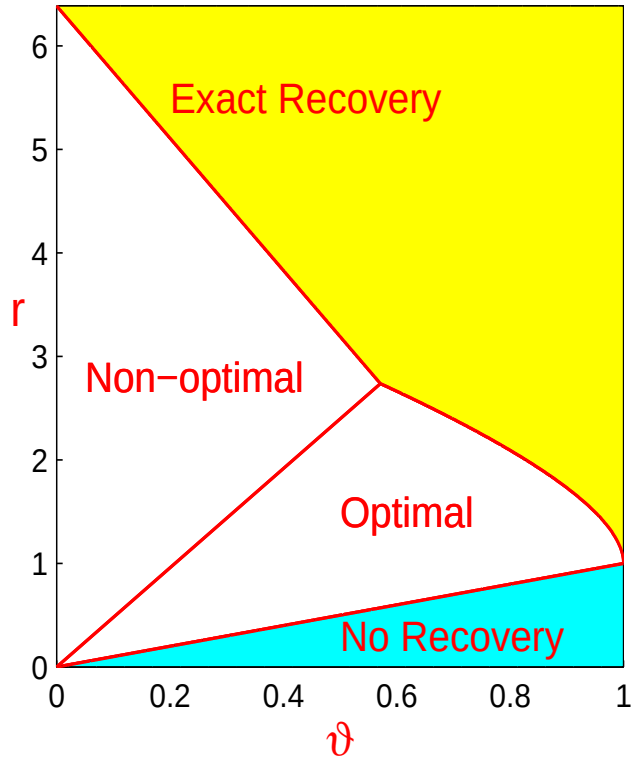
Method	L^0/L^1-penalization	CASE
Regime	Rare/Strong	Rare/Weak
Loss	$P\{S(\hat{\beta} = S(\beta)\}$ or L^q	$\text{Hamm}(\text{sgn}(\hat{\beta}), \text{sgn}(\beta))$
Optimality	Not in Hamming Error	Yes
Motivation	Imaging/Engineering	Genetics/Genomics
Design	Controllable/Nice	Uncontrollable/Bad
Key idea	One-stage global method	Multi-stage local method

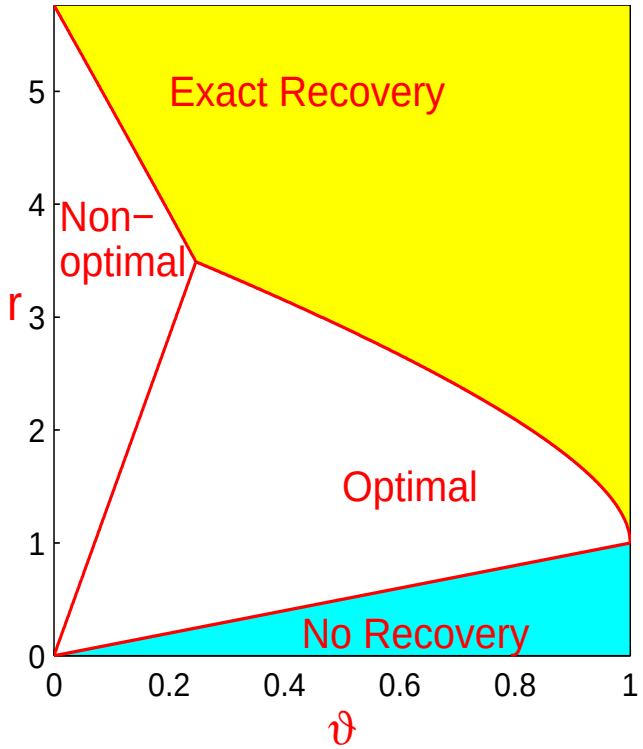
L^0/L^1 methods: Donoho and Stark (1989), Tibshirani (1996), and many others

Optimality of CASE ($X'X$ nearly tridiag)

$Y \sim N(X\beta, I_n)$, rows of X *iid* from $N(0, \frac{1}{n}\Omega)$, Ω is tri-diagonal







Optimality in more general designs: Ji and Jin (2012), Jin, Zhang and Zhnag (2014), Ke, Jin and Fan (2014)

Summary of Part IV

- Covariate-Assisted Ranking overcomes signal cancellation and provides more satisfying ranking than marginal ranking
- Post-screening subgraph consists of many small-size components, which can be cleaned separately
- CASE is optimal in terms of Hamming selection error, while L^0-L^1 methods are not

Main References

Genovese C, Jin J, Wasserman L and Yao Z (2012) A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107-2143.

Ji P and Jin J (2012) UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40**(1), 73-103.

Jin J, Zhang C-H and Zhang Q (2014) Optimality of Graphlet Screening in High Dimensional Variable Selection. *J. Mach. Learn. Res.* **15** 2723-2772

Ke Z, Jin J and Fan J (2014) Covariance assisted screening and estimation. *Ann. Statist.* **42**(6) 2202-2242.

Ke Z and Yang F (2017) Covariate assisted variable ranking. *arXiv.1705.10370*.