Clustering by Important Features PCA (IF-PCA)

Rare/Weak Signals and Phase Diagrams

Jiashun Jin, CMU

David Donoho (Stanford)
Zheng Tracy Ke (Univ. of Chicago)
Wanjie Wang (Univ. of Pennsylvania)

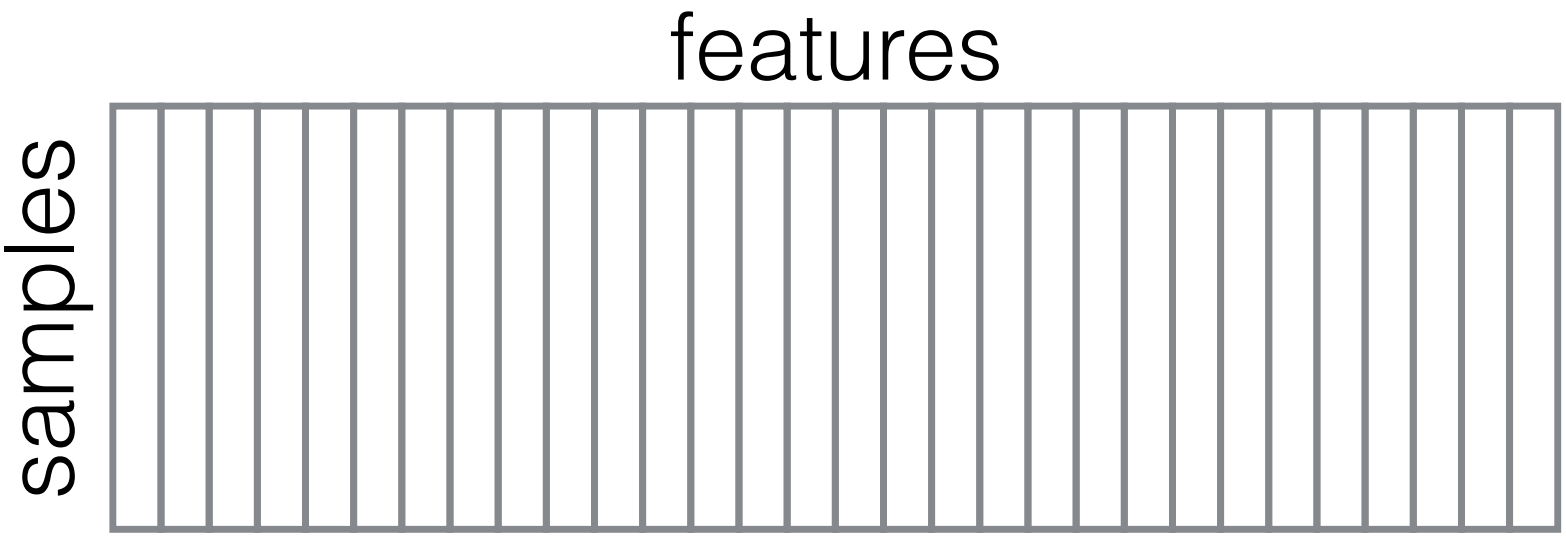
August 12, 2015

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 class labels

Left/right singular vectors



- Left singular vector:

  (n-dimensional) eigenvector of XX'

- Right singular vector:

  (p-dimensional) eigenvector of $X'X$

Principal Component Analysis (PCA)

Karl Pearson (1857-1936)

Idea:

- Transformation
- Dimension Reduction while keeping main info.
- Data = signal + noise (signal matrix: low-rank)

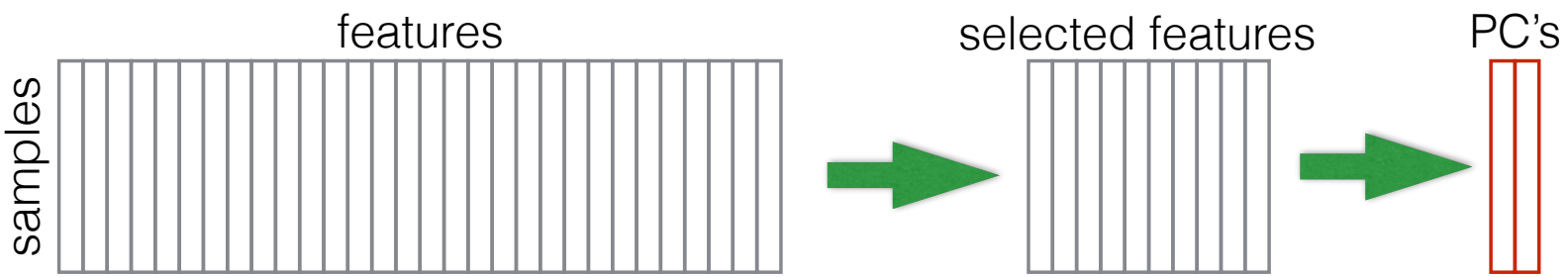
Microarray data (after standardization):

many columns of the signal matrix are 0

Important Features PCA (IF-PCA)

Idea: PCA applied to a small fraction of carefully selected features:

- Rank features by Kolmogorov-Smirnov statistic
- Select those with the largest KS-scores
- Apply PCA to the post-selection data matrix



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

IF-PCA (microarray data)

$$W_i(j) = [X_i(j) - \bar{X}(j)]/\hat{\sigma}(j): \quad \text{feature-wise normalization}$$

$$W = [w_1, \ldots, w_p] = [W_1', \ldots, W_n']', \qquad F_{n,j}(t) = \frac{1}{n}\sum_{i=1}^{n} 1\{W_i(j) \le t\}$$

1. Rank features with Kolmogorov-Smirnov (KS) scores

   $$\psi_{n,j} = \sqrt{n} \cdot \sup_{-\infty<t<\infty} |F_{n,j}(t) - \Phi(t)|, \qquad (\Phi\text{: CDF of } N(0,1))$$

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. Fix $t > 0$. $\hat{U}^{(t)} \in \mathbb{R}^{n,K-1}$: first $(K-1)$ **left** singular vectors of **post-selection data matrix** $[w_j : \psi^*_{n,j} \ge t]$

4. Apply classical k-means algorithm to $\hat{U}^{(t)} \in \mathbb{R}^{n,K-1}$

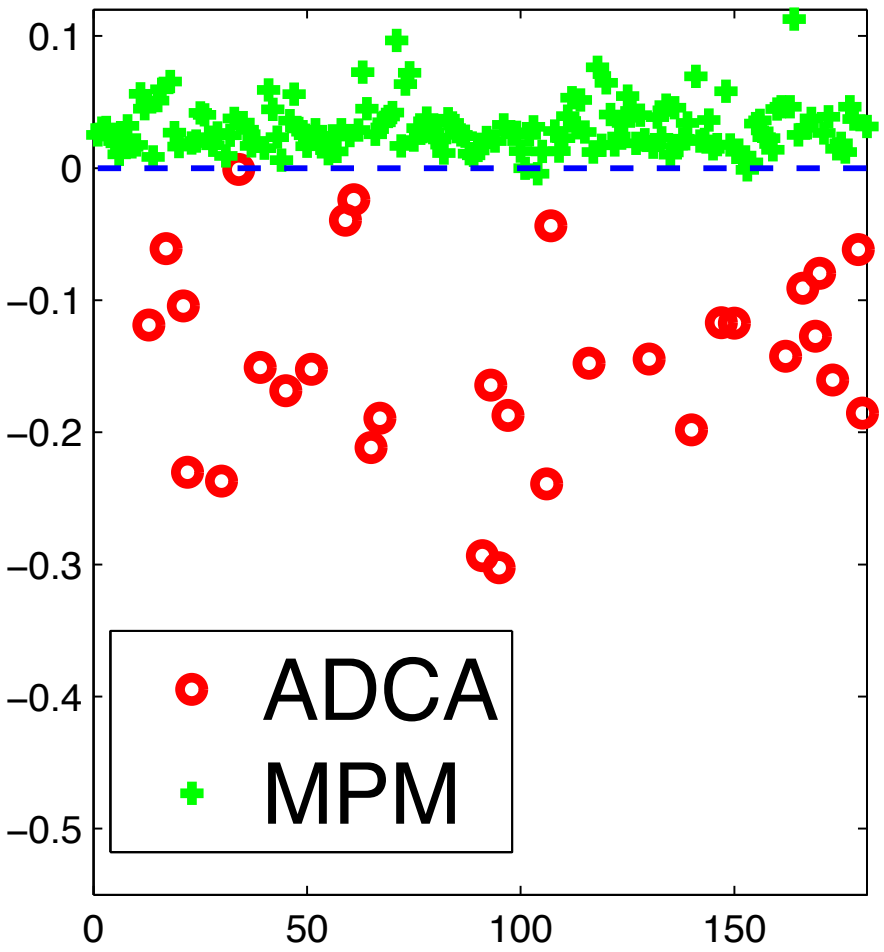
IF-PCA-HCT: $t = t_p^{HC}$: **H**igher **C**riticism threshold (**TBA**)

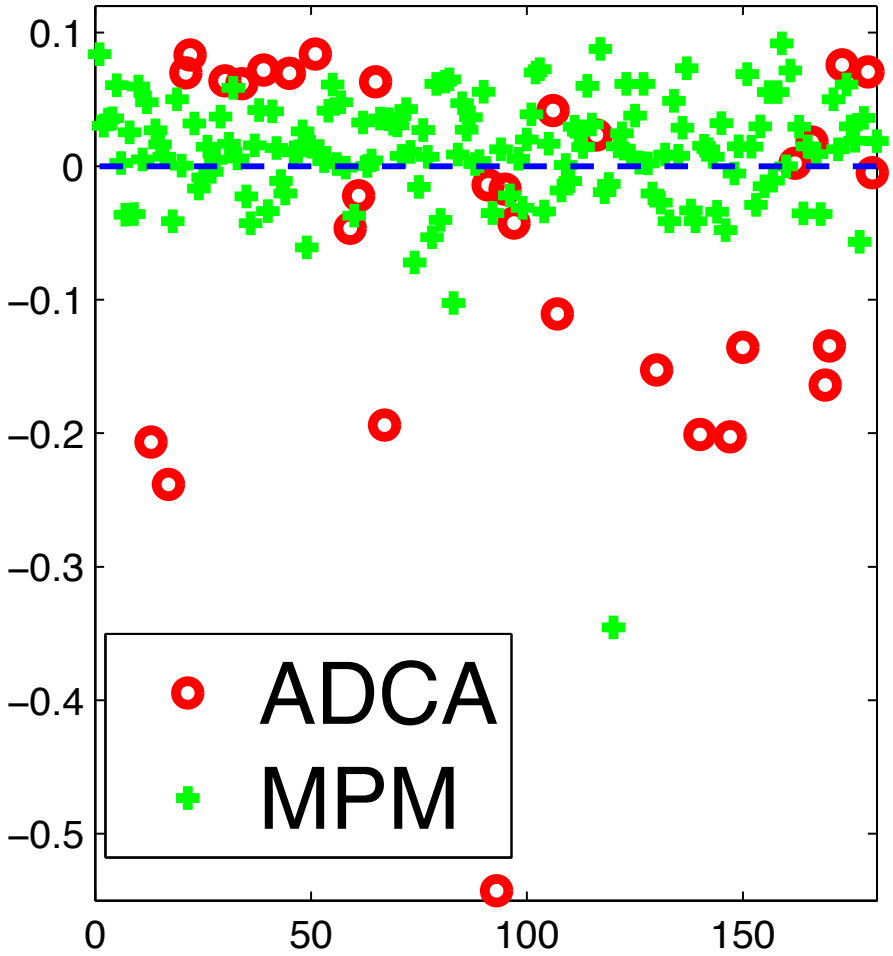
The blessing of feature selection

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

Left: plot of $\hat{U}_{HC}$ (Lung Cancer; $K = 2$ so $\hat{U}_{HC} \in \mathbb{R}^n$ is a vector)

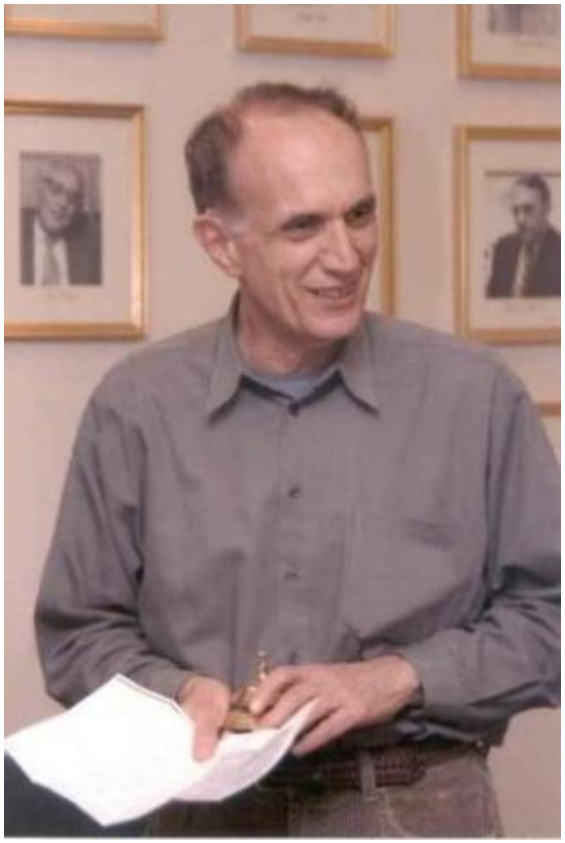
Right: counterpart of $\hat{U}_{HC}$ without feature selection

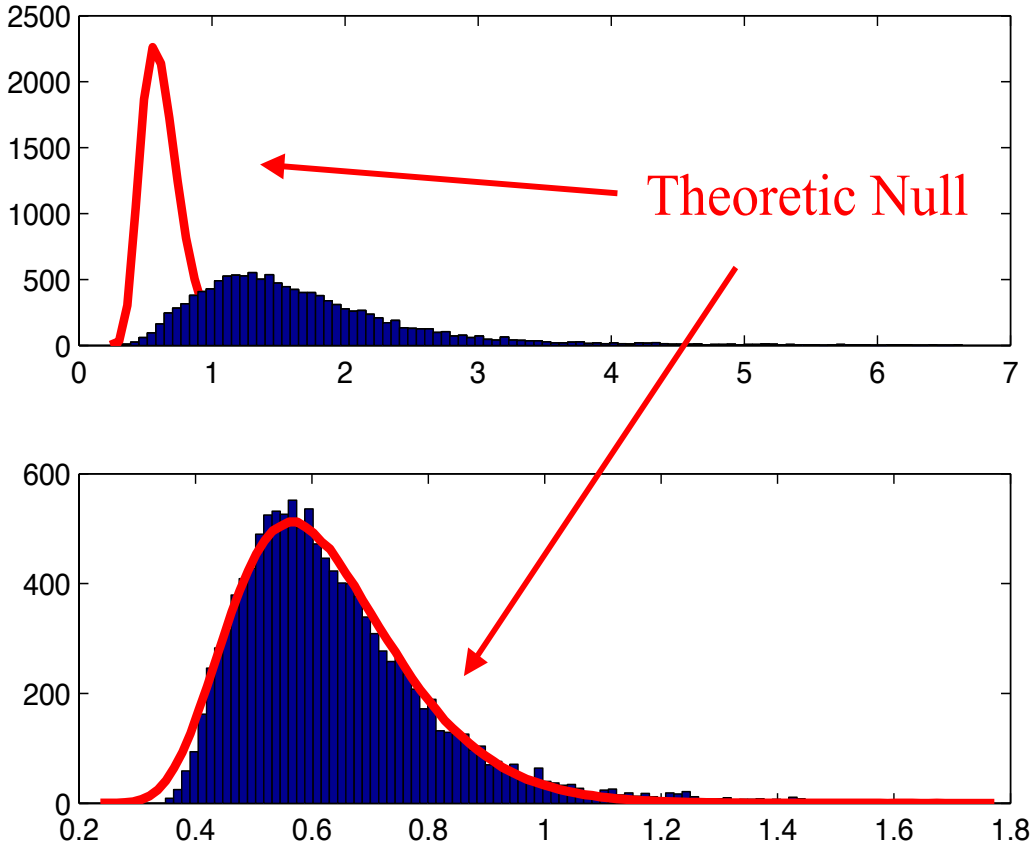




Efron's null correction (Lung Cancer)

Efron's theoretic Null: density of $\psi_{n,j}$ if $X_i(j) \overset{iid}{\sim} N(u_j, \sigma_j^2)$ (not depend on (j, u_j, σ_j); **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 directly implementable (class labels unknown)
- FDR: need tuning and target on **Rare/Strong** signals *[Benjaminin 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

Tukey's Higher Criticism

John W. Tukey (1915-2002)

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 --

Higher Criticism (HC)

Review papers: Donoho and Jin (2015), Jin and Ke (2015)

- Proposed by Donoho and Jin (2004) for sparse signal detection
- Found useful in GWAS, DNA Copy Number Variants (CNV), Cosmology and Astronomy, Disease surveillance
- Extended to many different directions: Innovated HC (for colored noise), signals detection in a regression model, HC when noise dist. is unknown/nonGaussian, estimate the proportion of non-null effects
- Threshold choice in classification [Donoho and Jin (2008, 2009), Fan et al (2013), Jin (2009)]

Threshold choice by HC for IF-PCA

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

$$t_p^{HC} = \mathrm{argmax}_t\{HC_p(t)\},$$

$$HC_p(t) = \frac{\sqrt{p}[\hat{G}_p(t) - \bar{F}_0(t)]}{\sqrt{\sqrt{n}[\hat{G}_p(t) - \bar{F}_0(t)] + \hat{G}_p(t)}} \qquad \textbf{(new)}$$

- $\bar{F}_0$: survival function of Efron's theoretical null
- $\hat{G}_p$: empirical survival function of renormalized KS-scores $\psi^*_{n,j}$

HCT: implementation

- Compute P-values: $\pi_j = \bar{F}_0(\psi^*_{n,j})$, $1 \le j \le p$
- Sort P-values: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$
- Define the HC functional by

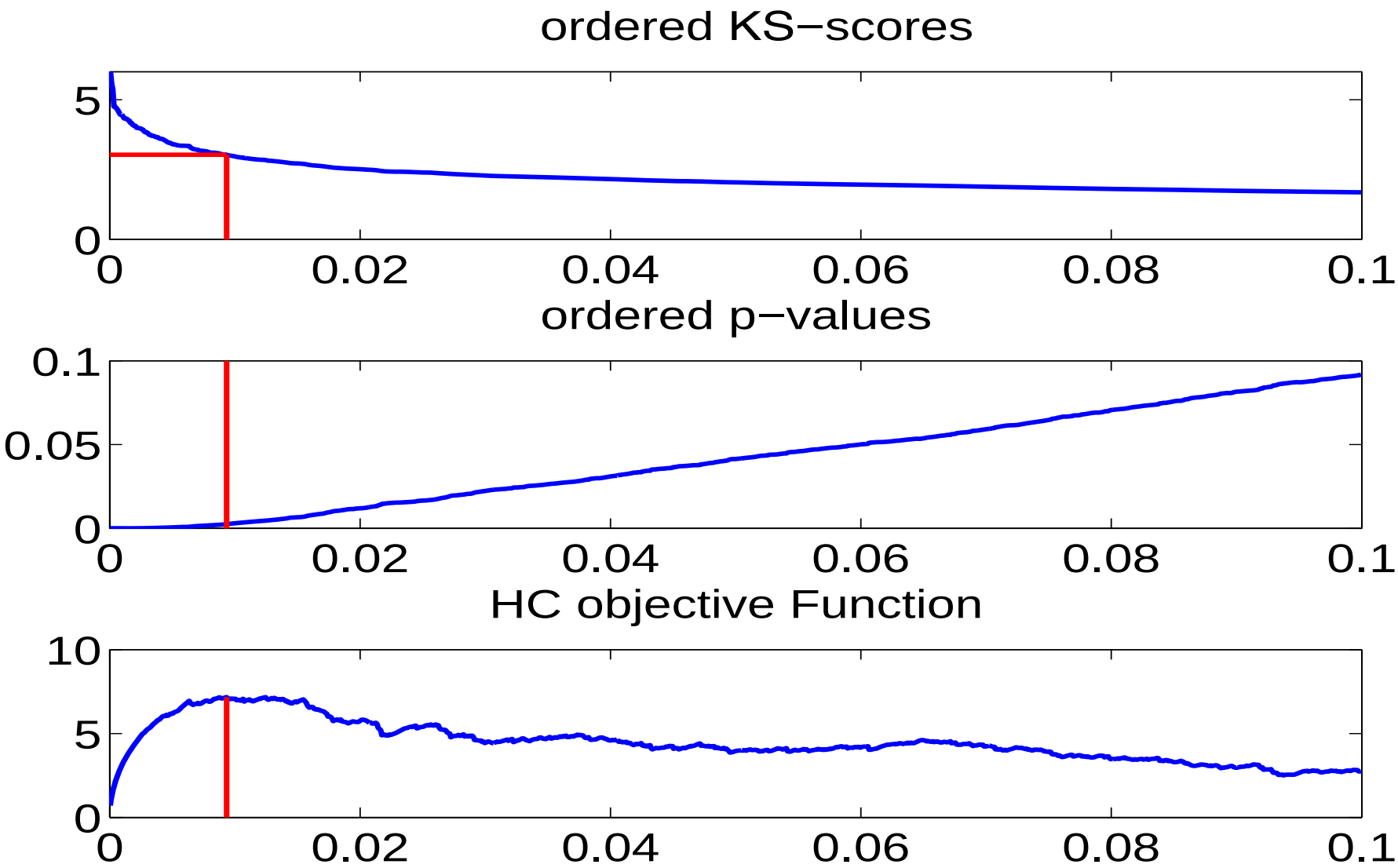
$$HC_{p,k} = \frac{\sqrt{p}\big(k/p - \pi_{(k)}\big)}{\sqrt{k/p + \max\{\sqrt{n}\big(k/p - \pi_{(k)}\big), 0\}}}$$

Let $\hat{k} = \mathrm{argmax}_{\{1 \le k \le p/2, \pi_{(k)} > \log(p)/p\}}\{HC_{p,k}\}$.

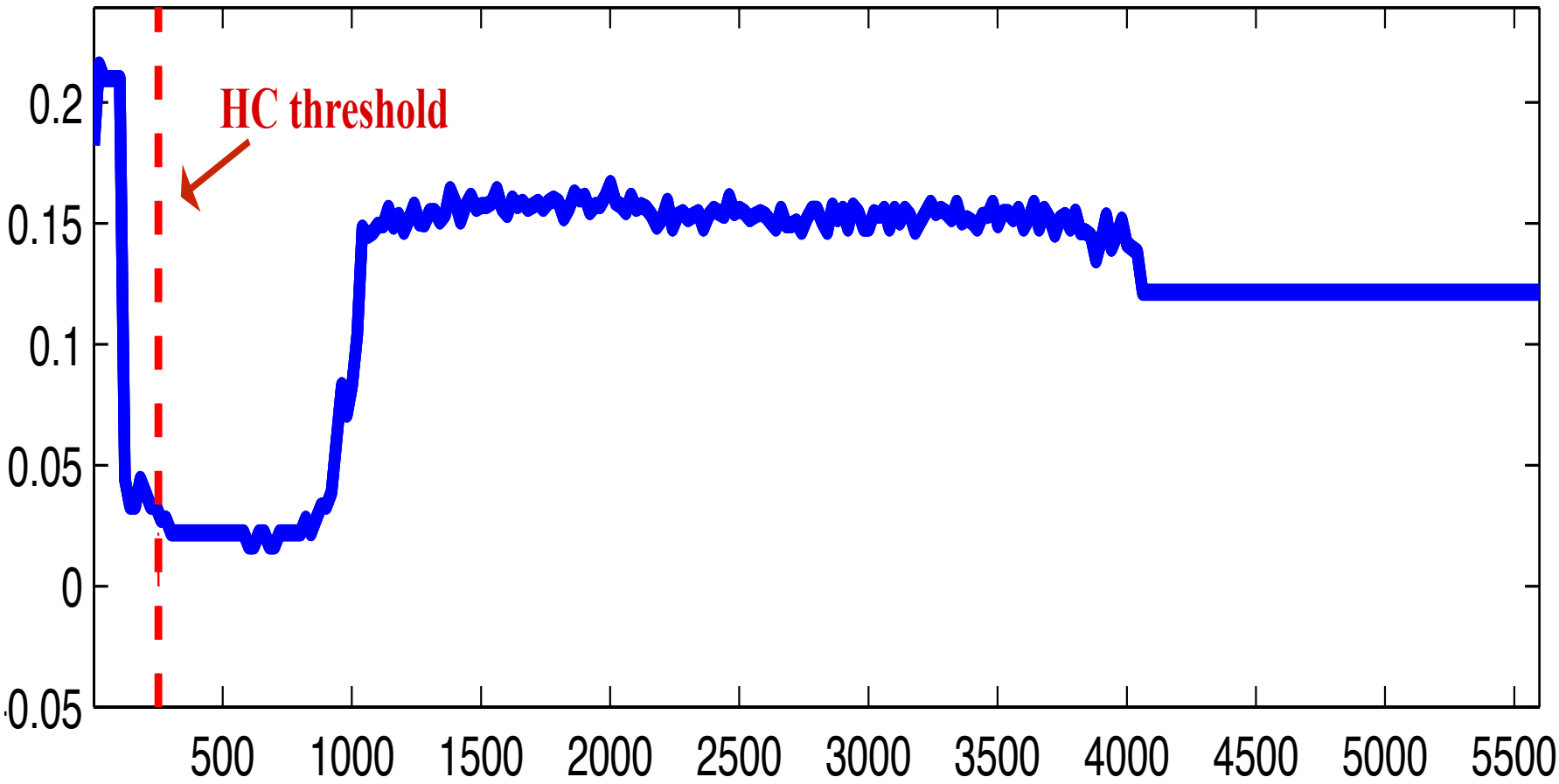
HC threshold t_p^{HC} is the $\hat{k}$-th largest KS-score

$$HC_p(t)\big|_{t=\psi^*_{(k)}} = \frac{\sqrt{p}[k/p - \pi_{(k)}]}{\sqrt{k/p + \sqrt{n}\big(k/p - \pi_{(k)}\big)}}; \quad \psi^*_{(k)}\text{: sorted KS scores}$$

Illustration



Illustration, II (Lung Cancer)



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

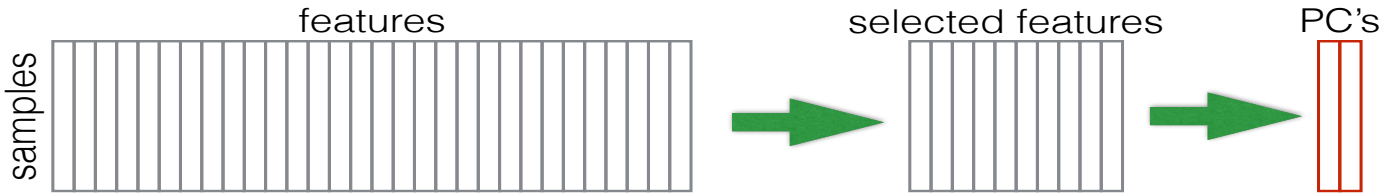
Comparison

$$r = \frac{\text{error rate of IF-PCA-HCT}}{\text{minimum error rate of all other methods}}$$

#	Data set	K	kmeans	kmeans++	Hier	SpecGem*	IF-PCA-HCT	r
1	Brain	5	.286	.427(.09)	.524	.143	.262	1.83
2	Breast Cancer	2	.442	.430(.05)	.500	.438	.406	.94
3	Colon Cancer	2	.443	.460(.07)	.387	.484	.403	1.04
4	Leukemia	2	.278	.257(.09)	.278	.292	.069	.27
5	Lung Cancer	2	.116	.196(.09)	.177	.122	.033	.29
6	Lung Cancer(2)	2	.436	.439(.00)	.301	.434	.217	.72
7	Lymphoma	3	.387	.317(.13)	.468	.226	.065	.29
8	Prostate Cancer	2	.422	.432(.01)	.480	.422	.382	.91
9	SRBCT	4	.556	.524(.06)	.540	.508	.444	.87
10	SuCancer	2	.477	.459(.05)	.448	.489	.333	.74

**: SpecGem is classical PCA [Lee et al (2010)]; Arthur and Vassilvitskii (2007)*

Sparse PCA and variants of IF-PCA



#	Data set	K	Clu-sPCA *	IF-kmeans	IF-Hier	IF-PCA-HCT
1	Brain	5	.172	.302	.476	.262
2	Breast Cancer	2	.438	.378	.351	.406
3	Colon Cancer	2	.404	.396	.371	.403
4	Leukemia	2	.292	.114	.250	.069
5	Lung Cancer	2	.110	.180	.177	.033
6	Lung Cancer(2)	2	.434	.226	.227	.217
7	Lymphoma	3	.055	.138	.355	.065
8	Prostate Cancer	2	.422	.382	.412	.382
9	SRBCT	4	.428	.417	.603	.444
10	SuCancer	2	.466	.430	.500	.333

$*$: project to estimated feature space (sparse PCA) and then clustering; unclear how to set λ (ideal λ is used above); clustering $\neq$ feature estimation

Zou et al (2006), Witten and Tibshirani (2010)

Summary (so far)

- IF-PCA-HCT consists of three simple steps
  - Marginal screening (KS)
  - Threshold choice (Empirical null + HCT)
  - Post-selection PCA
- tuning free, fast, and yet effective
- easily extendable and adaptable

Next: theoretical reasons for HCT in RW settings

RW viewpoint

In many types of "Big Data", signals of interest are not only **sparse (rare)** *but also individually* **weak**, *and we have no priori where these RW signals are*

- "Large p small n" (e.g., genetics and genomics)

$$(\text{Signal strength})^{\alpha} \quad \propto \quad n \quad \propto \quad \$ \text{ or labor}$$

  Clustering: $\alpha = 6$; classification: $\alpha = 2$

- Technical limitation (e.g., astronomy)
- Early detection (e.g., disease surveillance)

RW model and Phase Diagram

Many methods/theory target on **Rare/Strong** *signals* (*if conditions XX hold and all signals are sufficiently strong ...*)

Our proposal:

- RW model: parsimonious model capturing the main factors (sparsity and signal strength)
- Phase Diagram:
  - provides sharp results that characterize when the desired goal is **impossible** or **possible** to achieve
  - an approach to distinguish **non-optimal** and **optimal** procedures

Sparse signal detection (global testing)

Donoho and Jin (2004), Ingster (1997, 1999), Jin (2004)

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

$$H_0^{(p)}: \ \mu = 0, \quad \text{vs.} \quad H_1^{(p)}: \ \mu(j) \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}$$

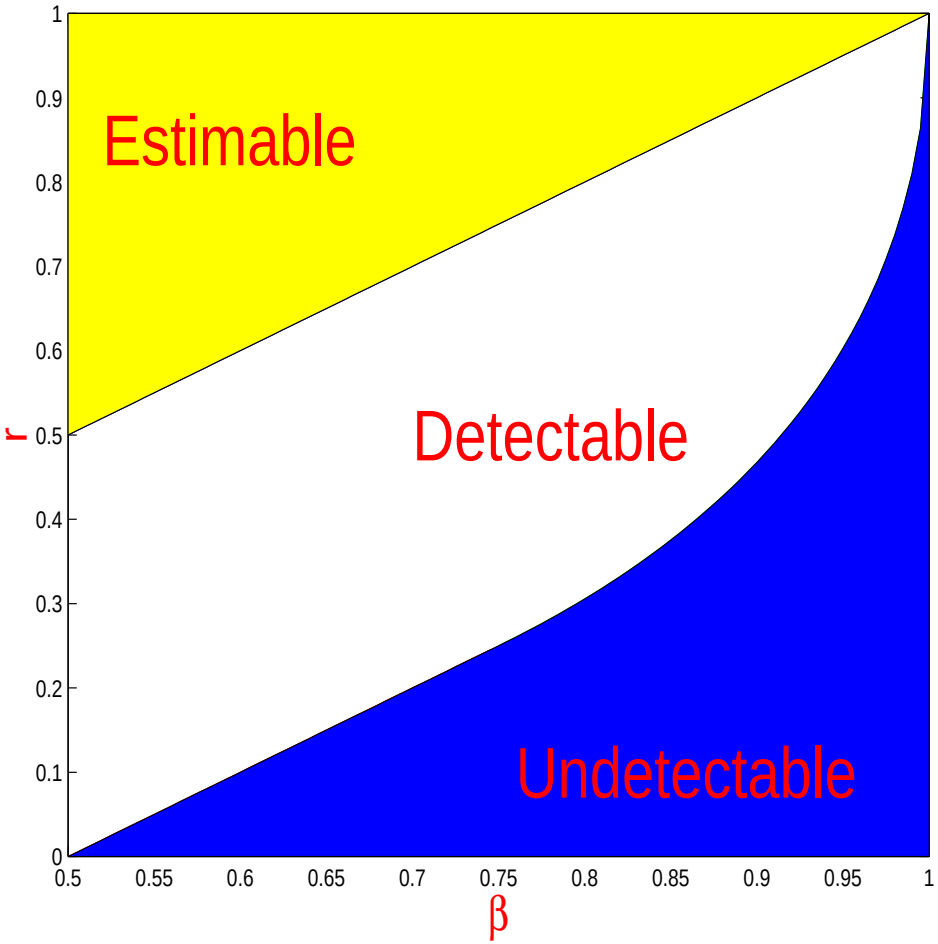
Calibration and subtlety of the problem:

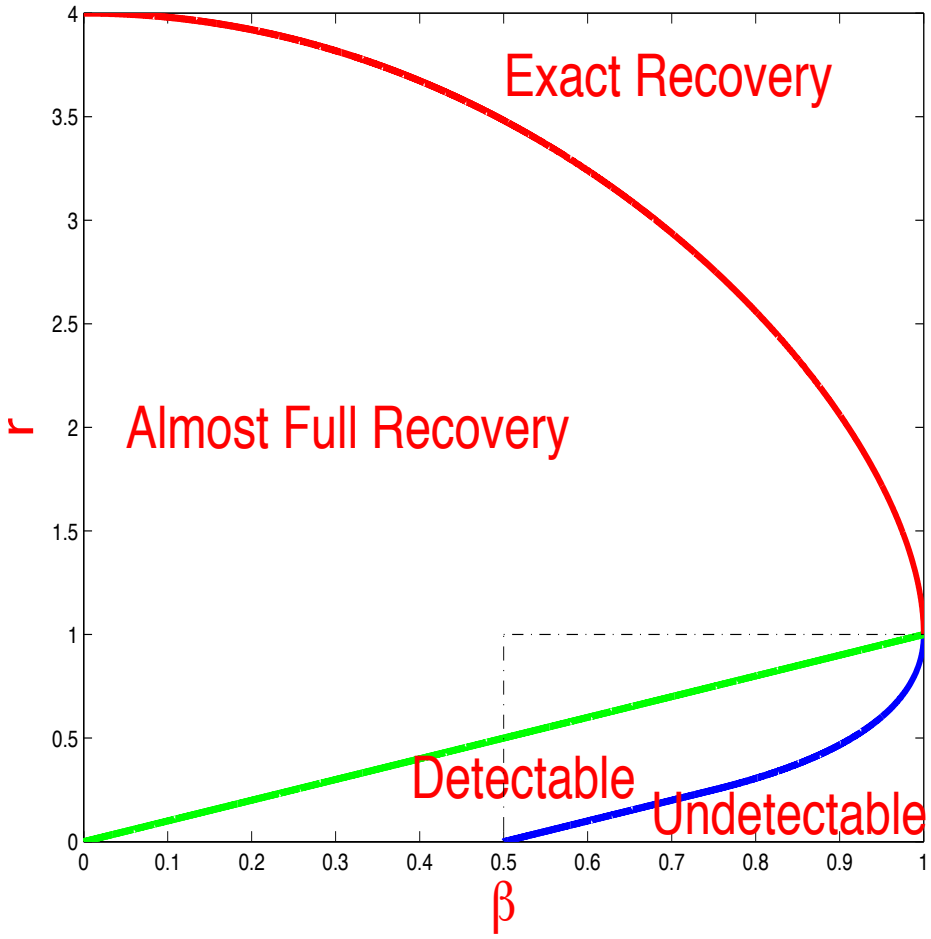
$$\epsilon_p = p^{-\beta}, \quad \tau_p = \sqrt{2r\log(p)}, \qquad 0 < \beta < 1, \ r > 0$$

- Rare: **only a small** fraction of non-zero means
- Weak: signals **only moderately** strong

Phase Diagram (signal detection/recovery)

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





RW settings: why we care?

- Growing awareness of irreproducibility
- Many methods/theory focus on Rare/Strong signals, do not cope well with RW settings
  - FDR-controlling methods (showed before)
  - L^0/L^1-penalization methods
  - Minimax framework

Ioannidis, (2005). "Why most published research findings are false";

Donoho and Stark (1989); Chen et al (1995); Tibshirani (1996);

Abrmovich et al (2006); Jin et al (2015); Ke et al (2015)

A two-class model for clustering

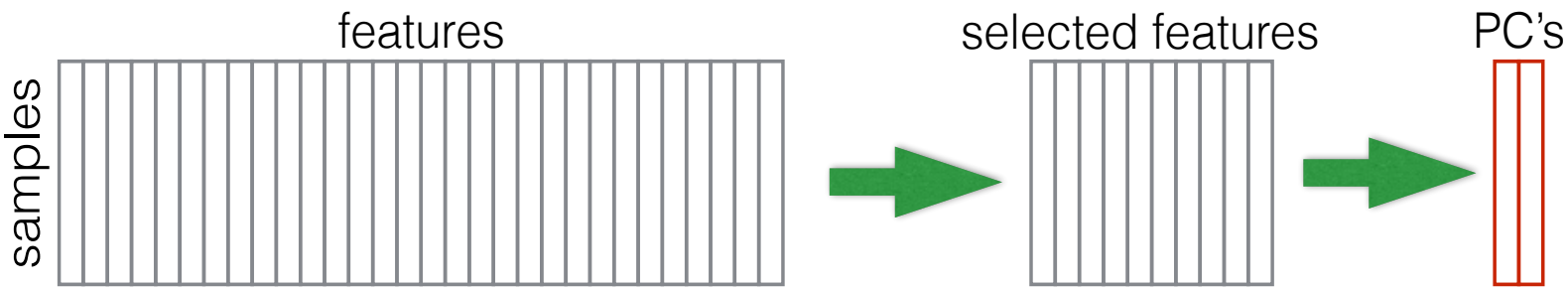
Jin, Ke & Wang (2015)

$$X_i = \ell_i \mu + Z_i, \quad Z_i \overset{iid}{\sim} N(0, I_p), \quad i = 1, \ldots, n \quad (p \gg n)$$

- $\ell_i = \pm 1$: unknown class labels (**main interest**)
- $\mu \in R^p$: feature vector
- RW: only a small fraction of entries of μ is nonzero, each contributes weakly to clustering

Goal. Theoretical insight on HCT (and more)

IF-PCA simplified to two-class model



	microarray	two-class model
pre-normalization	yes	**skipped**
feature-wise screening	Kolmogorov-Smirnov	**chi-square**
	$\psi_j = \sup_t \lvert F_{n,j}(t) - \Phi(t)\rvert$	$\psi_j = (\lVert x_j\rVert - n)/\sqrt{2n}$
re-normalization	Efron's null correction	**skipped**
threshold choice	HCT	same
post-selection PCA	same	same

Asymptotic Rare/Weak (ARW) model

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

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

$$\mu(j) \overset{\textit{iid}}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}$$

- "large p small n":

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

- RW signal:

$$\epsilon_p = p^{-\beta}, \qquad \tau_p = \sqrt[4]{(4/n) r \log(p)}$$

Note: $\psi_j = (\|x_j\|^2 - n)/\sqrt{2n} \approx N(0,1)$ *or* $N(\sqrt{2r\log(p)}, 1)$

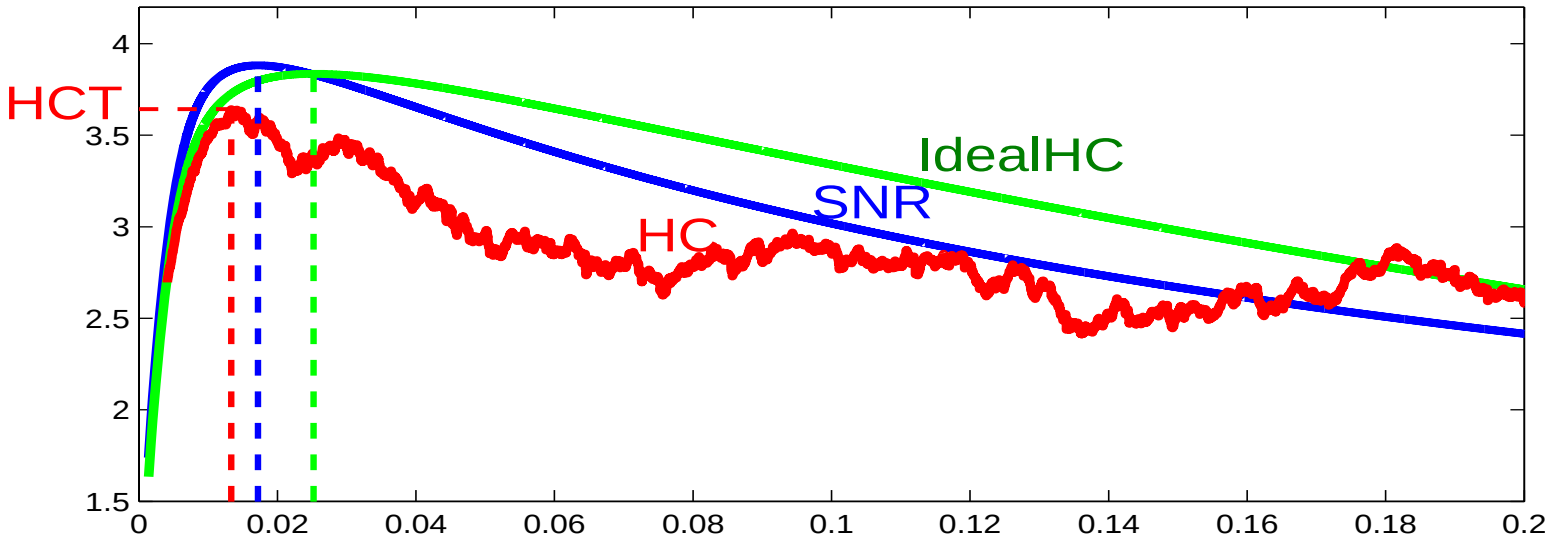
Three functions: *HC*, *idealHC*, $\widetilde{SNR}$

Given $X = \ell\mu' + Z$, we retain feature j if $\psi_j \geq t$

- $\bar{F}_0(t)$: survival function of normalized $\chi^2_{2n}(0)$
- $\hat{G}_p(t)$: empirical survival function ψ_j
- $\hat{U}^{(t)}$: first left singular vector of $[x_j : \psi_j \geq t]$

1. $HC(t)$:
   $\sqrt{p}[\hat{G}_p(t) - \bar{F}_0(t)]/[\sqrt{n}[\hat{G}_p(t) - \bar{F}_0(t)] + \hat{G}_p(t)]^{\frac{1}{2}}$
2. $idealHC(t)$:
   $HC(t)$ with $\hat{G}_p(t)$ replaced by $\bar{G}_p(t)$
3. $\widetilde{SNR}(t)$:
   $\hat{U}^{(t)} \propto \widetilde{SNR}(t) \cdot \ell + z + rem,\ z \sim N(0, I_n)$

Three thresholds: $t_p^{HC} \approx t_p^{idealHC} \approx t_p^{ideal}$



$$\widetilde{SNR}(t) = \frac{E[\|\mu^{(t)}\|^2]}{\sqrt{E[\|\mu^{(t)}\|^2] + E[|\hat{S}(t)|]/n}} \propto \frac{\sqrt{p}(\bar{G}_p(t) - \bar{F}_0(t))}{\sqrt{\sqrt{n}[\bar{G}_p(t) - \bar{F}_0(t)] + \bar{G}_p(t)}}$$

- $\mu^{(t)}$: μ restricted to $\hat{S}(t)$ (index set of all retained features)
- $E\|\mu^{(t)}\|^2 \approx \tau_p^2 p[\bar{G}_p(t) - \bar{F}_0(t)]$
- $E\|\mu^{(t)}\|^2 + E[|\hat{S}(t)|/n] \propto \tau_p^2[\bar{G}_p(t) - \bar{F}_0(t)] + \bar{G}_p(t)/n$

Impossibility

Let $\rho(\beta)$ be the standard phase function (before). Define

$$\rho_\theta(\beta) = (1-\theta)\rho(\frac{1}{2} + \frac{\beta - \frac{1}{2}}{1-\theta}); \qquad \text{recalling } n = p^\theta$$

Consider IF-PCA with $t > 0$. Let $\hat{U}^{(t)} \in \mathbb{R}^n$ be the first left singular vector of post-selection data matrix

$$[x_j : \psi_j \geq t]$$

Consider ARW with

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

If $r < \rho_\theta(\beta)$, then for any threshold t,

$$\mathrm{Cos}(\hat{U}^{(t)}, \ell) \leq c_0 < 1 \qquad \text{(IF-PCA partially fails)}$$

Possibility

If $r > \rho_\theta(\beta)$, then

- For some t, $\mathrm{Cos}(\hat{U}^{(t)}, \ell) \to 1$
- There is a non-stochastic function $\widetilde{SNR}(t)$ such that for some t, $\widetilde{SNR}(t) \gg 1$ and

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

- HCT yields the right threshold choice:

$$t_p^{HC}/t_p^{ideal} \to 1 \text{ in prob.}; \quad t_p^{ideal} = \mathrm{argmax}_t\{\widetilde{SNR}(t)\}$$

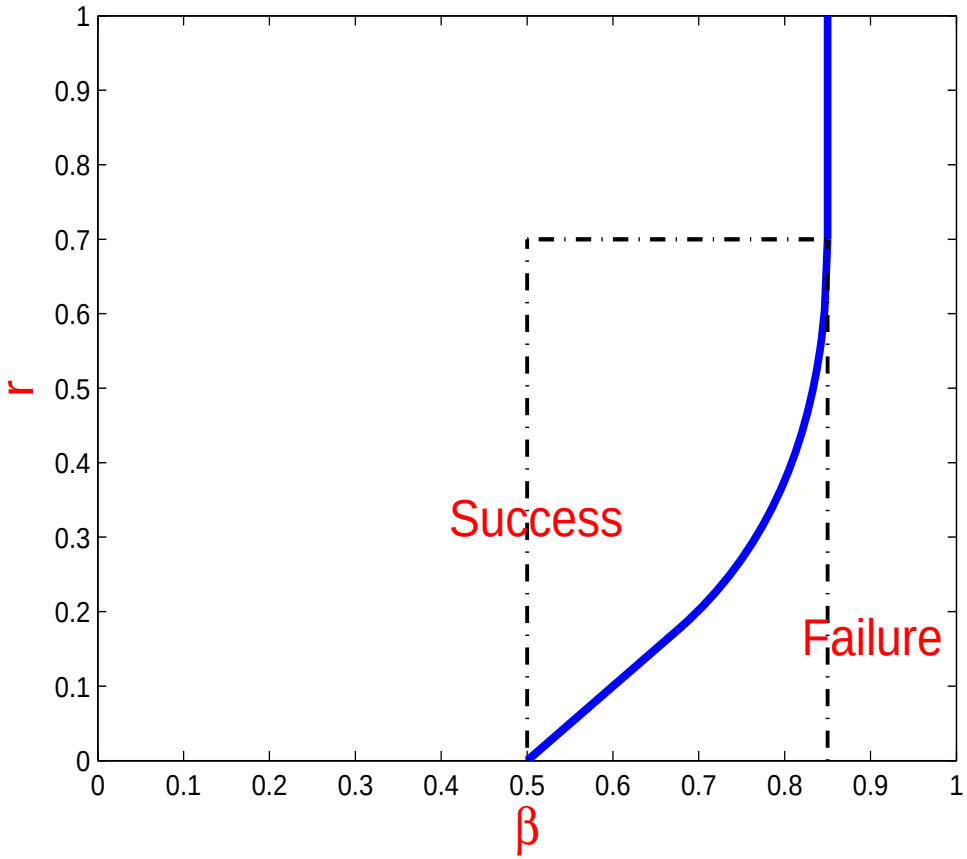
- IF-PCA-HCT yields successful clustering:

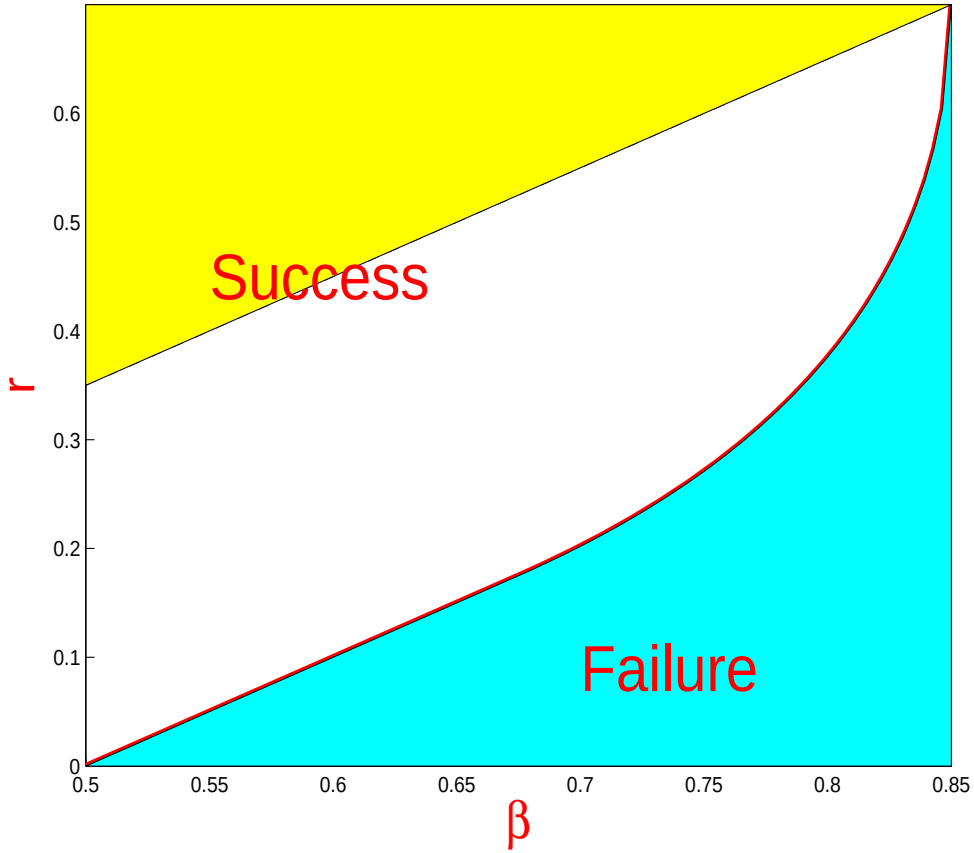
$$\mathrm{Hamm}_p(\hat{\ell}^{HCT}; \beta, r, \theta) \to 0 \qquad \dagger$$

$\dagger$: $\mathrm{Hamm}_p(\hat{\ell}; \beta, \alpha, \theta) = (n/2)^{-1} \min_{b=\pm\mathrm{sgn}(\hat{\ell})} \{\sum_{i=1}^n P(b_i \neq \mathrm{sgn}(\ell_i))\}$

Phase Diagram (IF-PCA)

$\#(\text{useful features}) \approx p^{1-\beta}$, $\tau_p = \sqrt[4]{(4/n)r\log(p)}$; $n = p^{\theta}$ $(\theta = .6)$





Summary (so far)

- Big Data are here, but signals are RW
- RW model and Phase Diagrams: theoretical framework specifically for RW settings, allow for insights that will be otherwise overlooked
- HC provides the right threshold choice

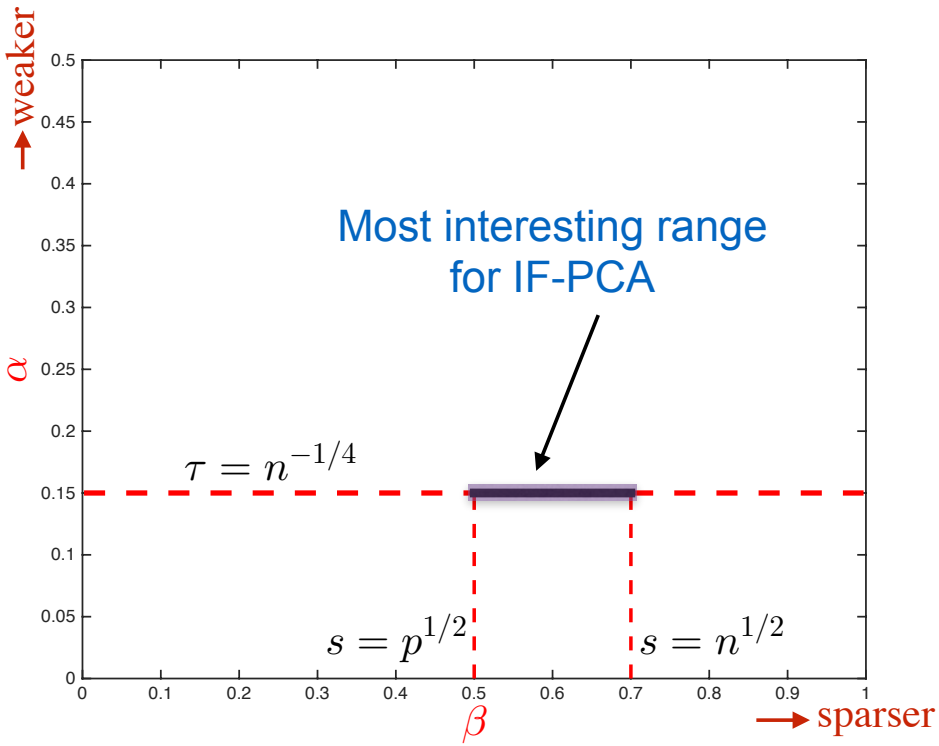
Limitation: IF-PCA *only* works at a **specific** signal strength and in a *limited* sparsity range

Want a more complete story:

statistical limits for clustering under ARW

ARW for statistical limits (a slight change)

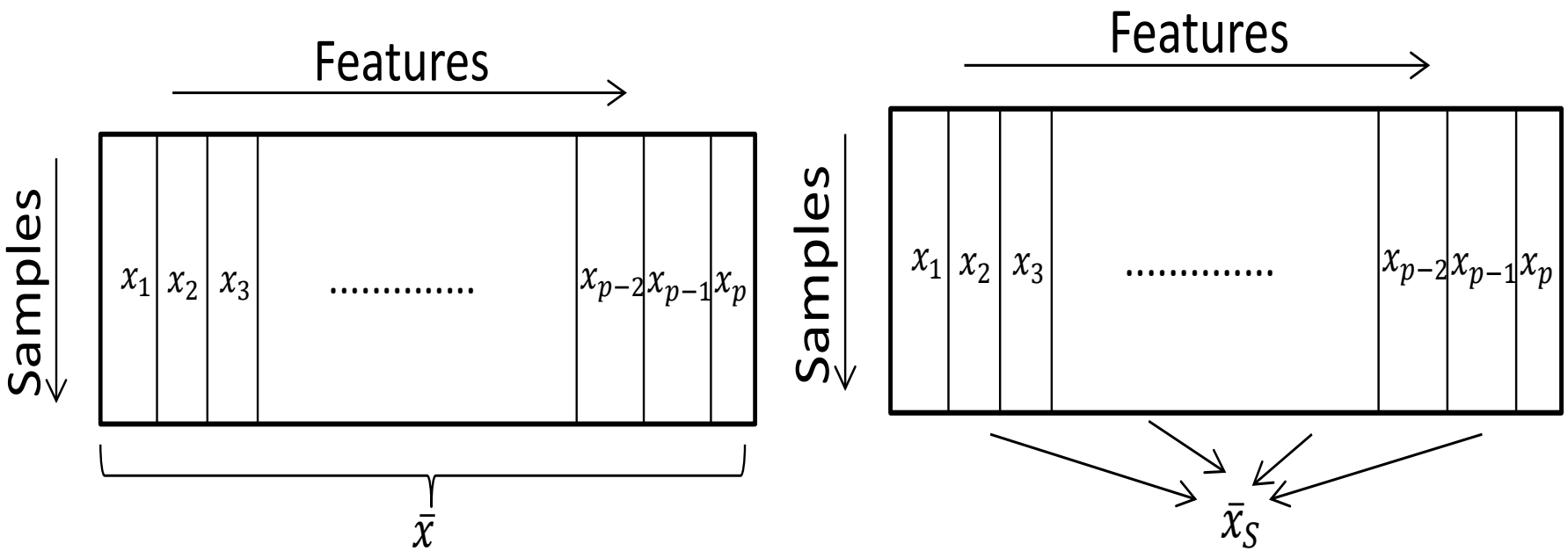
- $X = \ell\mu' + Z$
- $\ell_i = \pm 1$ with equal prob.
- $\mu(j) \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}$
- $n = p^\theta$, $\epsilon_p = p^{-\beta}$



	For statistical limits	For IF-PCA
τ_p	$\tau_p = p^{-\alpha}$, $\alpha > 0$	$\tau_p = \sqrt[4]{(4/n) r \log(p)}$
$s := \#\{\text{signals}\}$	$1 \ll s \ll p$	$\sqrt{n} \ll s \ll \sqrt{p}$

Aggregation methods

- Simple Aggregation: $\hat{\ell}_*^{sa} = \mathrm{sgn}\big(\bar{x}\big)$
- Sparse Aggregation: $\hat{\ell}_N^{sa} = \mathrm{sgn}\big(\bar{x}_{\hat{S}}\big)$, where $\hat{S} = \hat{S}(N) = \mathrm{argmax}_{\{S:|S|=N\}}\big\{\|\bar{x}_S\|_1\big\}$



Comparison of methods

Method	Simple Agg. $\hat{\ell}_*^{sa}$	PCA $\hat{\ell}_*^{if}$	Sparse Agg. $\hat{\ell}_N^{sa}$ $(N \ll p)$	IF-PCA $\hat{\ell}_t^{if}(t > 0)$
Sparsity	dense	dense	sparse	sparse
Strength	weak	weak	strong*	strong
F. Selection	No	No	Yes	Yes
Complexity	Poly.	Poly.	NP-hard	Poly.
Tuning	No	No	Yes	Yes**

- Notation. $\hat{\ell}_t^{if}$: IF-PCA adapted to two-class model
- Notation. $\hat{\ell}_*^{if}$: classical PCA (a special case)

*: signals are comparably stronger but still weak

**: a tuning-free version exists

Statistical limits (clustering)

$$\mathrm{Hamm}_p(\hat{\ell};\beta,\alpha,\theta) = (n/2)^{-1} \min_{b=\pm\mathrm{sgn}(\hat{\ell})} \Big\{ \sum_{i=1}^{n} P\big(b_i \neq \mathrm{sgn}(\ell_i)\big) \Big\}$$

$$\eta_\theta^{clu}(\beta) = \begin{cases} (1-2\beta)/2, & 0 < \beta < \frac{1-\theta}{2} \\ \theta/2, & \frac{1-\theta}{2} < \beta < 1-\theta \\ (1-\beta)/2, & \beta > 1-\theta \end{cases}$$

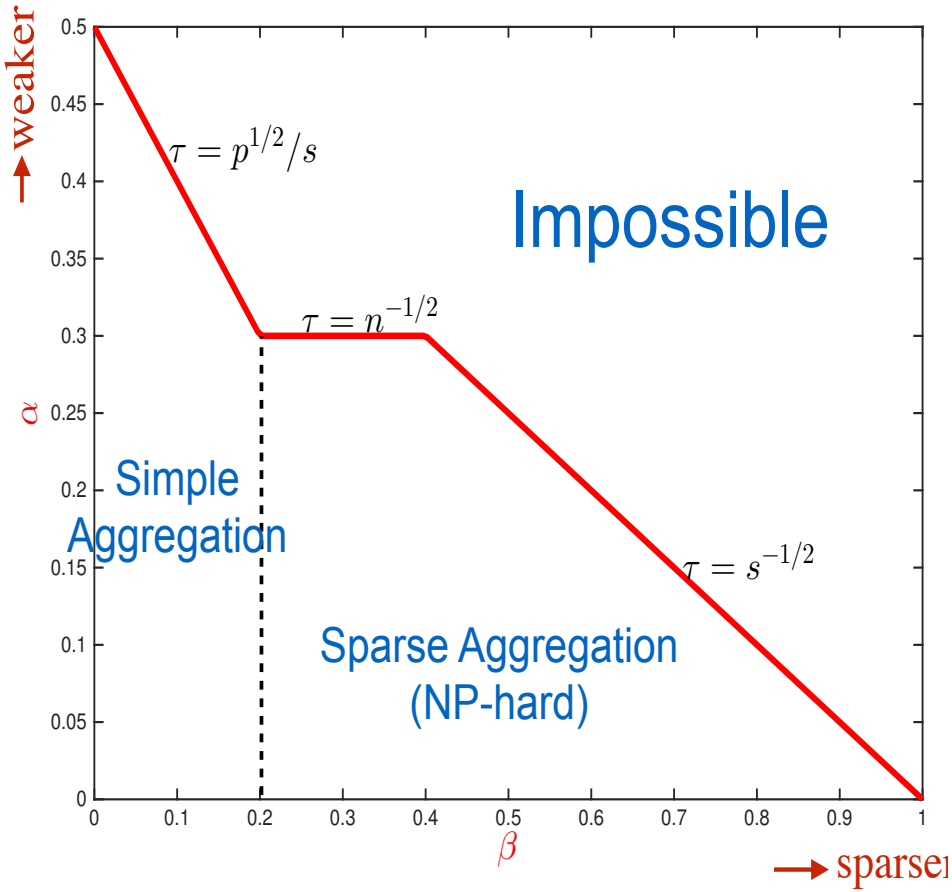
Consider ARW where

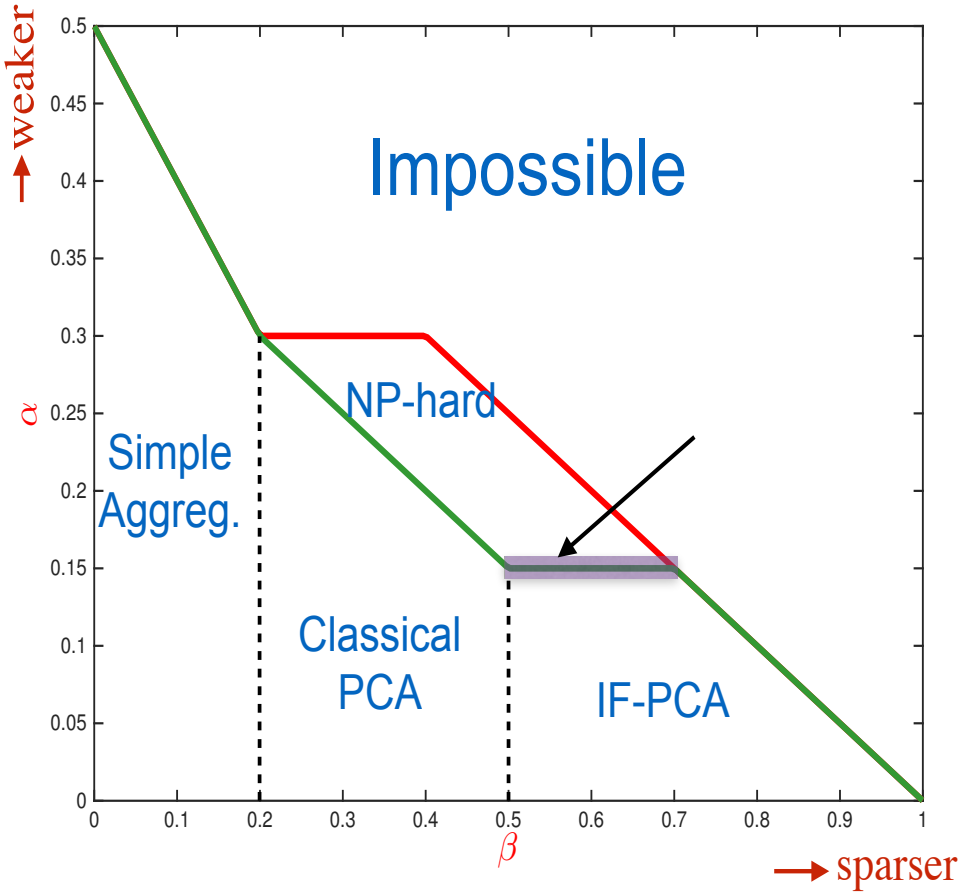
$$n = p^\theta, \qquad \epsilon_p = p^{-\beta}, \qquad \tau_p = p^{-\alpha}$$

- When $\alpha > \eta_\theta^{clu}(\beta)$, $\mathrm{Hamm}_p(\hat{\ell};\beta,\alpha,\theta) \gtrsim 1$
- When $\alpha < \eta_\theta^{clu}(\beta)$,
  - $\mathrm{Hamm}_p(\hat{\ell}_*^{sa};\beta,\alpha,\theta) \to 0$ for $\beta < \frac{1-\theta}{2}$;
  - $\mathrm{Hamm}_p(\hat{\ell}_N^{sa};\beta,\alpha,\theta) \to 0$ for $\beta > \frac{1-\theta}{2}$ $(N = p^{1-\beta})$

Phase Diagram (clustering; $\theta = 0.6$)

	For statistical limits	For IF-PCA
τ_p	$\tau_p = p^{-\alpha}$, $\alpha > 0$	$\tau_p = \sqrt[4]{(4/n) r \log(p)}$
Range of β	$0 < \beta < 1$	$1/2 < \beta < 1 - \theta/2$





Two closely related problems

$$X = \ell\mu' + Z, \qquad Z\text{: iid } N(0,1) \text{ entries}$$

- **(sig)**. Estimate support of μ (**Sig**nal recovery)

$$\mathrm{Hamm}_p(\hat{\mu}; \beta, r, \theta) = (p\epsilon_p)^{-1} \sum_{i=1}^{n} P(\mathrm{sgn}(\hat{\mu}_i) \neq \mathrm{sgn}(\mu_i))$$

- **(hyp)**. Test $H_0^{(p)}$ that $X = Z$ against alternative $H_1^{(p)}$ that $X = \ell\mu' + Z$ (global **hyp**othesis testing)

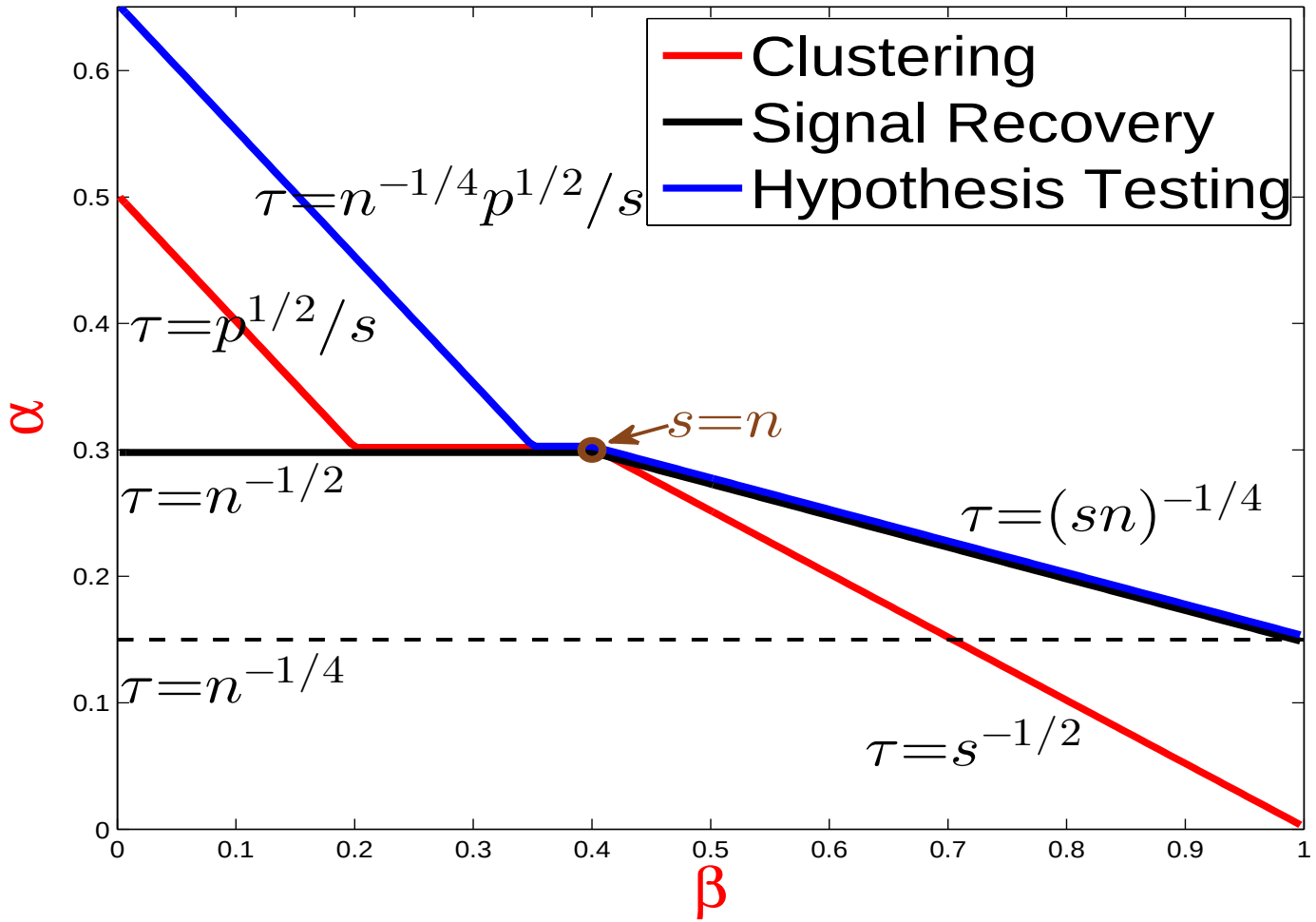
$$\mathit{TestErr}_p(\hat{T}, \beta, r, \beta) = P_{H_0^{(p)}}(\text{Reject } H_0) + P_{H_1^{(p)}}(\text{Accept } H_0^{(p)})$$

Arias-Castro & Verzelen (2015), Johnstone & Lu (2001), Rigollet & Berthet (2013)

Statistical limits (three problems)

$$\mu(j) \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p},$$

$n = p^\theta$, $s \equiv \#\{\text{useful features}\} \approx p^{1-\beta}$, signal strength $= p^{-\alpha}$

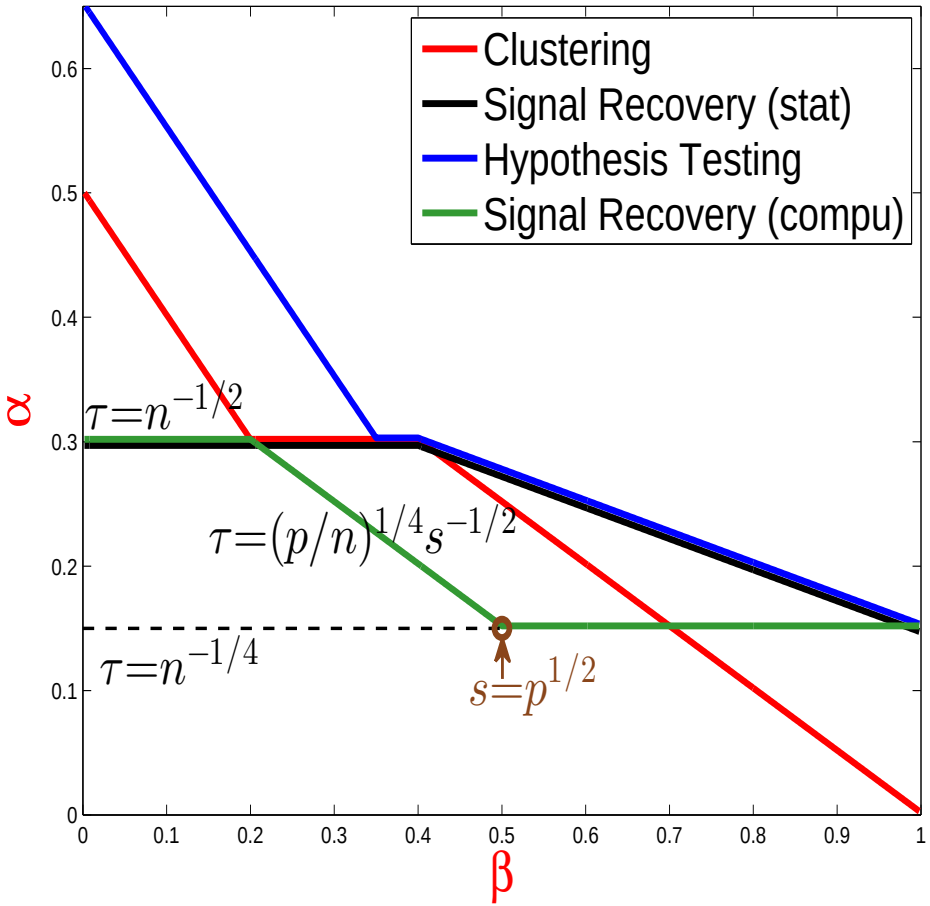


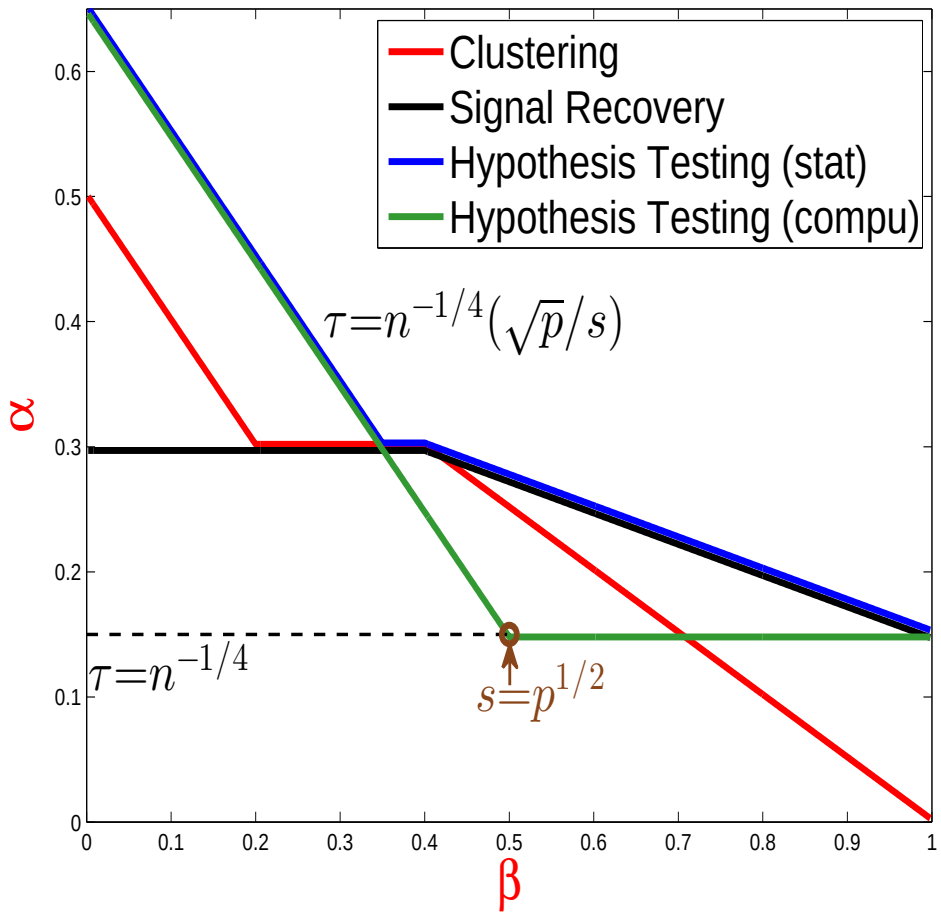
Computable upper bounds (two problems)

Left: signal recovery. Right: (global) hypothesis testing

$$\mu(j) \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p},$$

$n = p^\theta$, $s \equiv \#\{\text{useful features}\} \approx p^{1-\beta}$, signal strength $= p^{-\alpha}$





Take-home message

- "Big Data" is here, but your signals are RW; we need to do many things very differently
- RW model and Phase Diagrams are theoretical framework specifically designed for RW settings, and expose many insights we do not see with more traditional frameworks
- IF-PCA and Higher Criticism are simple and easy-to-adapt methods which are provably effective for analyzing real data, especially for Rare/Weak settings

Acknowledgements

Many thanks for *David Donoho, Stephen Fienberg, Peter Hall, Jon Wellner* and *many others* for inspirations, collaborations, and encouragements!

Main references.

Jin J, Wang W (2015) Important Features PCA for high dimensional clustering. *arXiv.1407.5241*.

Jin J, Ke Z, Wang W (2015) Phase transitions for high dimensional clustering and related problems. *arXiv.1502.06952*.

Two review papers on HC.

D. Donoho and J. Jin (2015). Higher Criticism for Large-Scale Inference, especially for rare and weak effects. *Statistical Science*, **30**(1), 1-25.

J. Jin and Z. Ke (2015). Rare and weak effects in large-scale inference: methods and phase diagrams. *arXiv.1410.4578*.