Clustering by Important Features PCA (IF-PCA)

Rare/Weak Signals and Phase Diagrams

Jiashun Jin, CMU

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

August 5, 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

Principal Component Analysis (PCA)

Karl Pearson (1857-1936)

Idea:

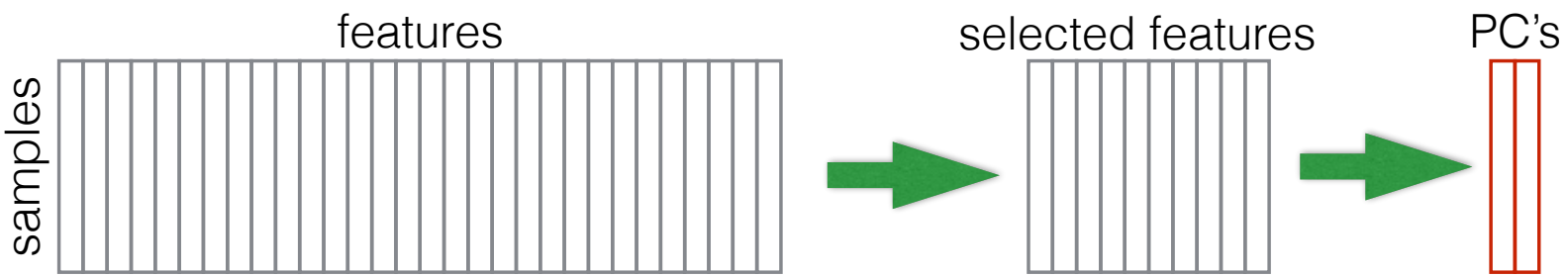
- Transformation
- Dimension Reduction while keeping main info.
- Data = signal + noise (signal matrix: low-rank)

**microarray data: many columns of the signal matrix are 0 after standardization*

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 & IF-PCA-HCT (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 a threshold $t > 0$. Let $\hat{U}^{(t)} \in \mathbb{R}^{n,K-1}$ be the first $(K-1)$ singular vectors of matrix $[w_j : \psi^*_{n,j} \ge t]$

   Our proposal : $t = t_p^{HC}$: **H**igher **C**riticism threshold (**TBA**),

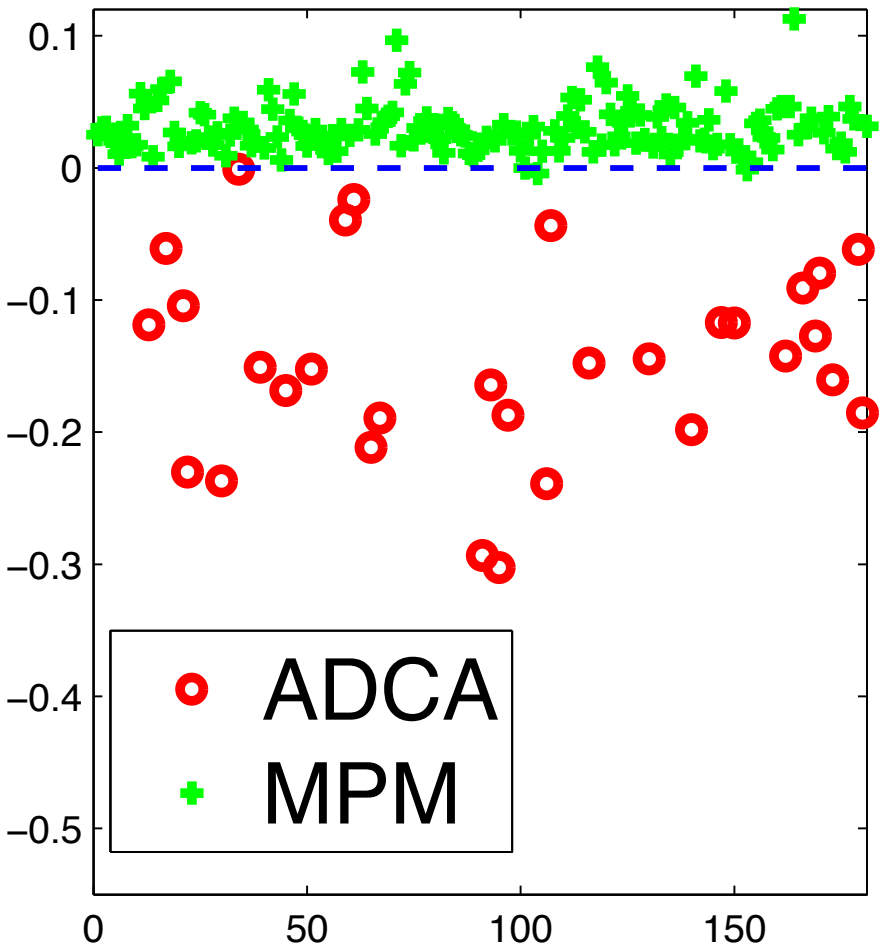
4. Apply classical k-means algorithm to $\hat{U}_{HC} \equiv \hat{U}^{(t)}\big|_{t=t_p^{HC}}$

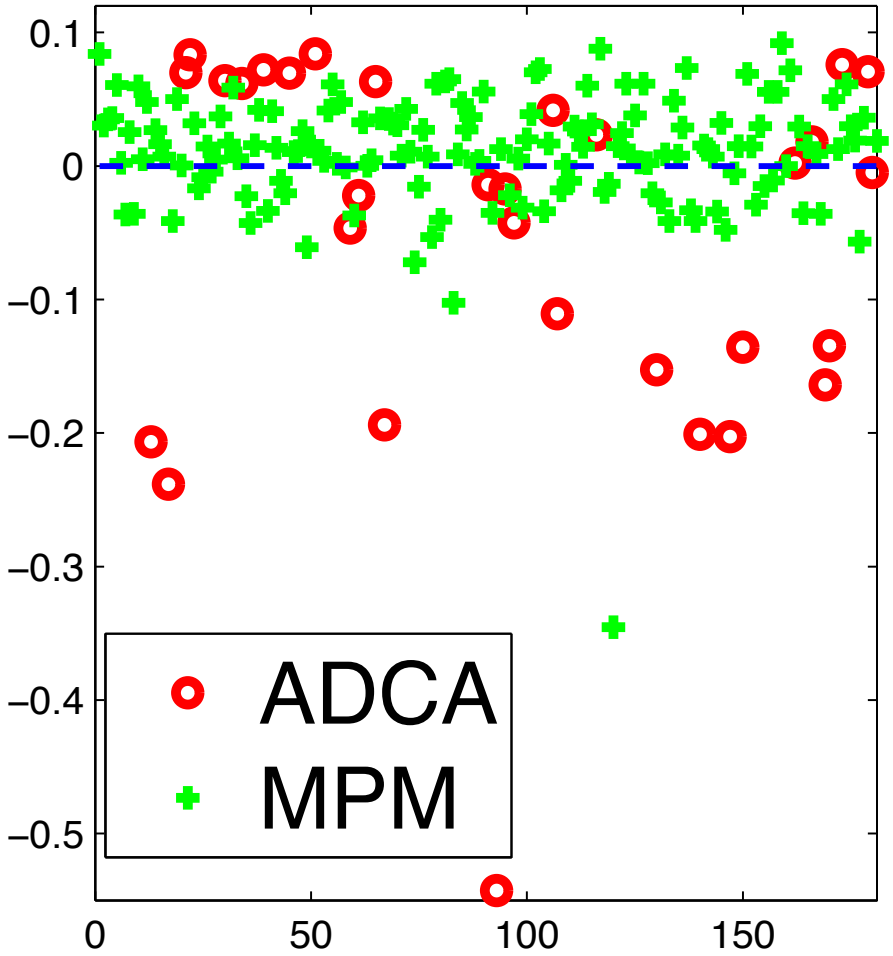
The blessing of feature selection

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

x-axis: $1, 2, \ldots, n$; y-axis: entries of $\hat{U}_{HC}$

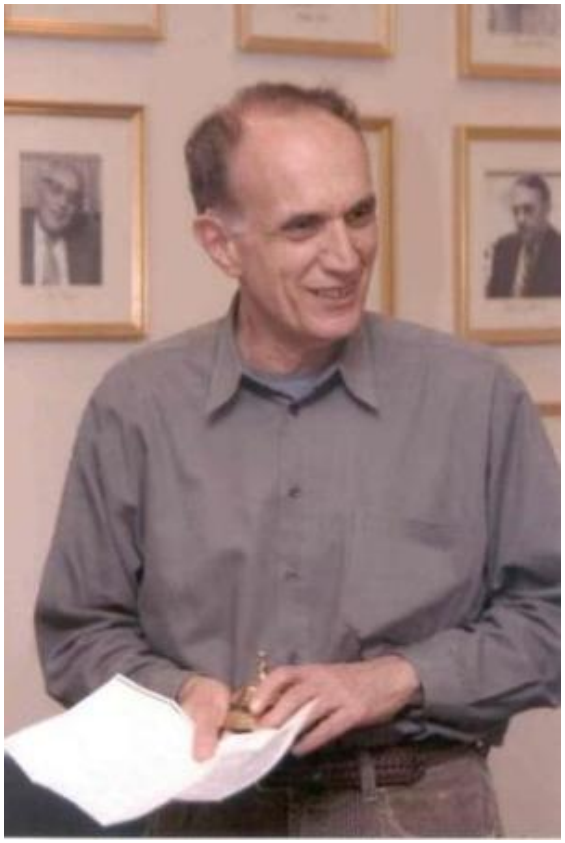
Right: counterpart of $\hat{U}_{HC}$ without feature selection

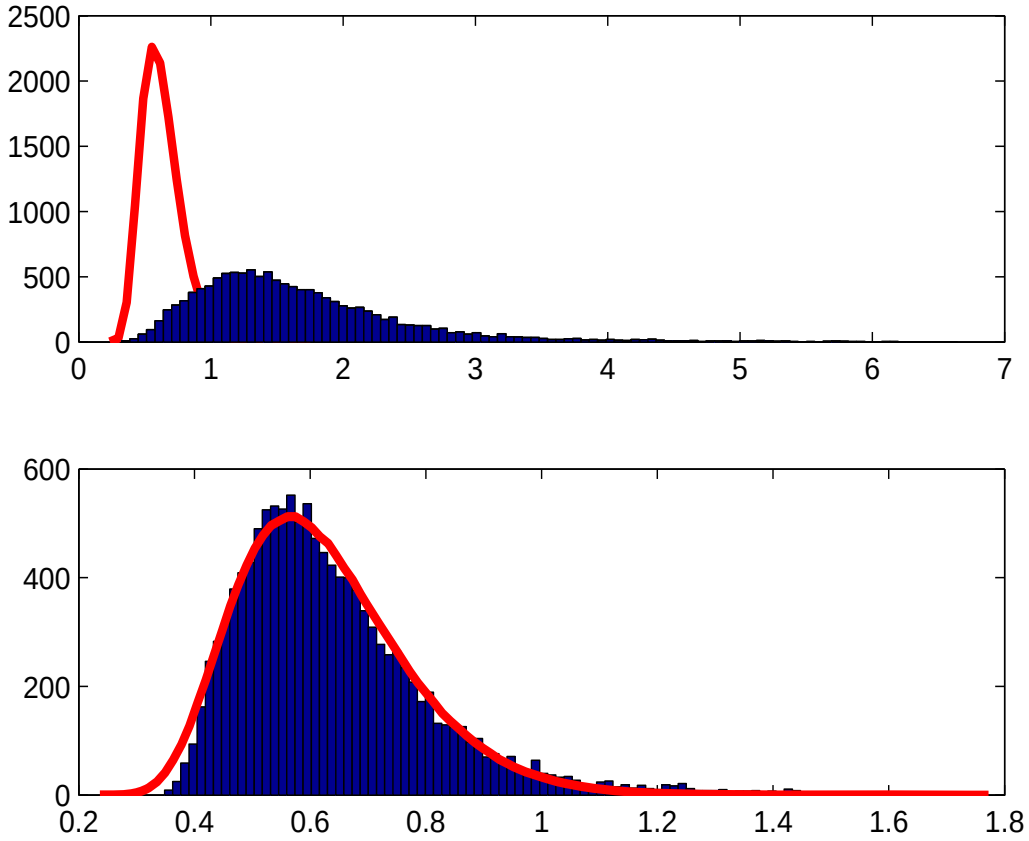




Efron's null correction (Lung Cancer)

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 (**red**) 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)]

Our proposal. Setting t by Higher Criticism (**RW** settings)

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)

- First proposed by Donoho and Jin (2004) for detecting sparse signals
- Also useful for setting threshold in cancer classification [Donoho and Jin (2008, 2009), Jin (2009)]
- Found useful in GWAS, DNA Copy Number Variants (CNV), Cosmology and Astronomy, Disease surveillance
- Extended to many different directions

Higher Criticism Threshold (HCT)

- Obtain P-values: $\pi_j = 1 - F_0(\psi^*_{n,j})$, $1 \le j \le p$

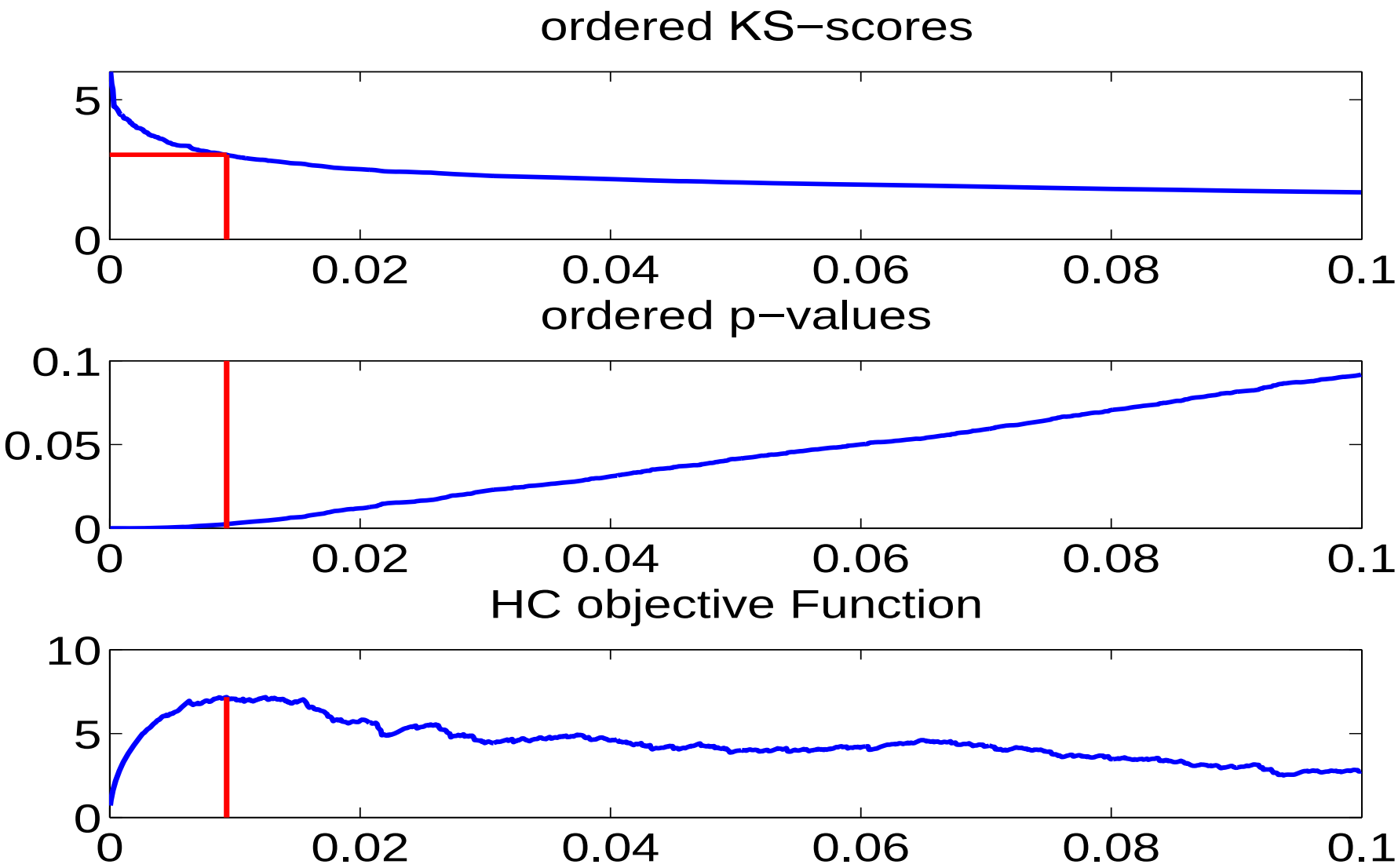
  [F_0: CDF of Efron's theoretical null]
- Sort P-values: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$
- Define the HC functional by

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

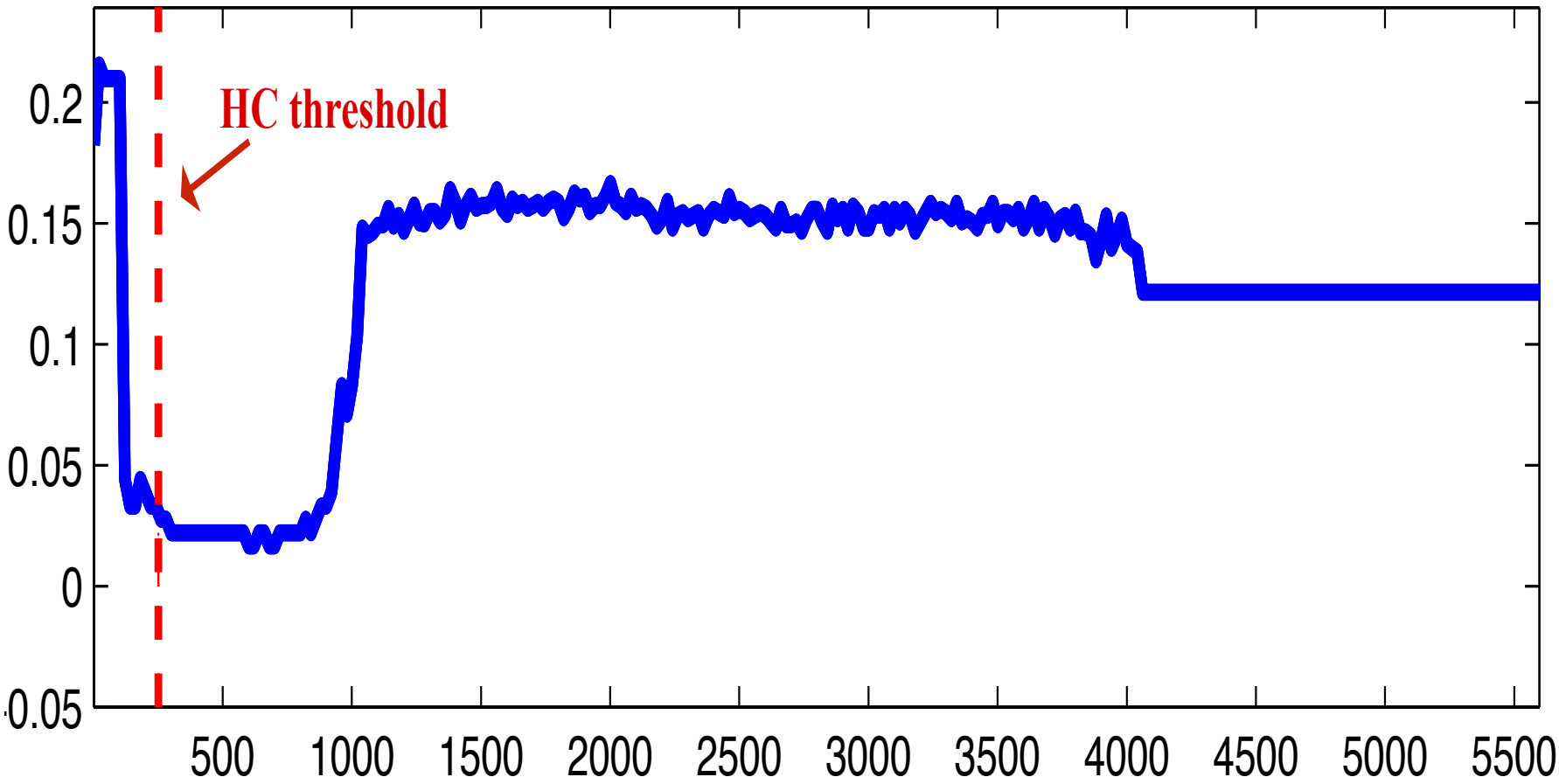
Let $\hat{k} = \mathrm{argmax}_{\{1 \le k \le p/2, \pi_{(k)} > \log(p)/p\}}\{HC_{p,k}\}$.

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

Illustration



Illustration, II (Lung Cancer)



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

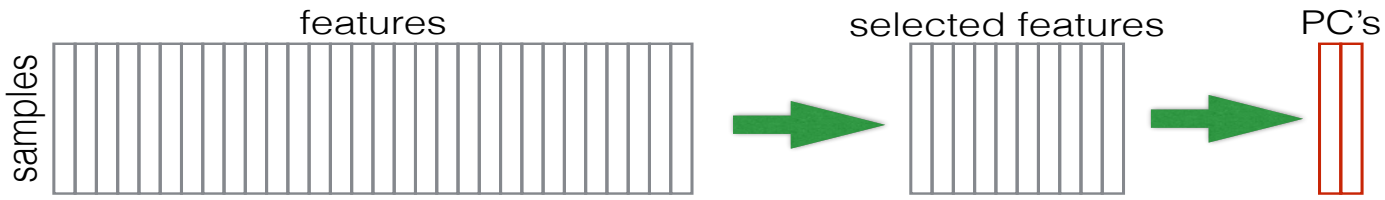
Comparison

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

#	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

Arthur and Vassilvitskii (2007), Hastie et al (2009), Lee et al (2010)

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

Remaining problems:

- Why HCT
- Statistical limits for clustering/related problems

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)

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

  Clustering: $\alpha = 4$ or 6, classification: $\alpha = 2$
- Technical limitation (e.g., astronomy)
- Early detection (e.g., disease surveillance)

RW model and Phase Diagram

Many scientific findings are not reproducible [Ioannidis, (2005). "Why most published research findings are false"] and 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

A two-class model for clustering

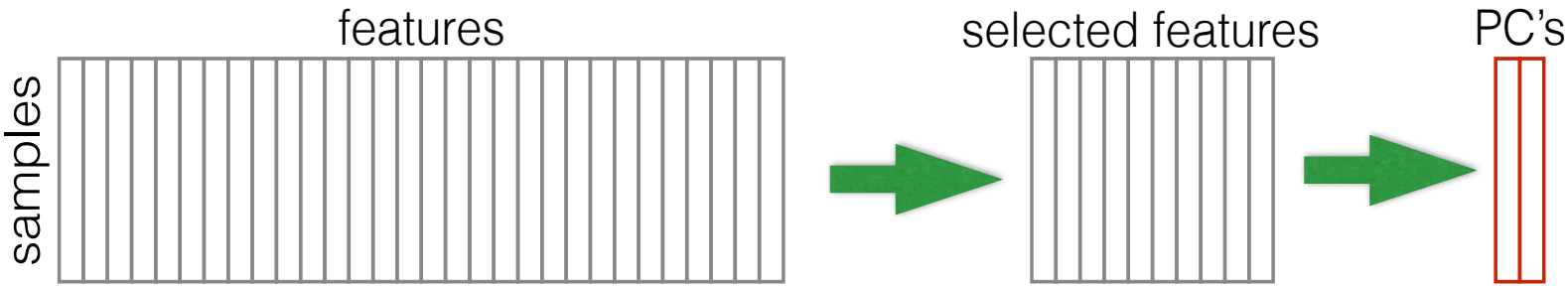
Jin, Ke & Wang (2015)

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

- $\ell_i = \pm 1$: 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

Interest. Rationale of HCT and Statistical limits

IF-PCA simplified to two-class model

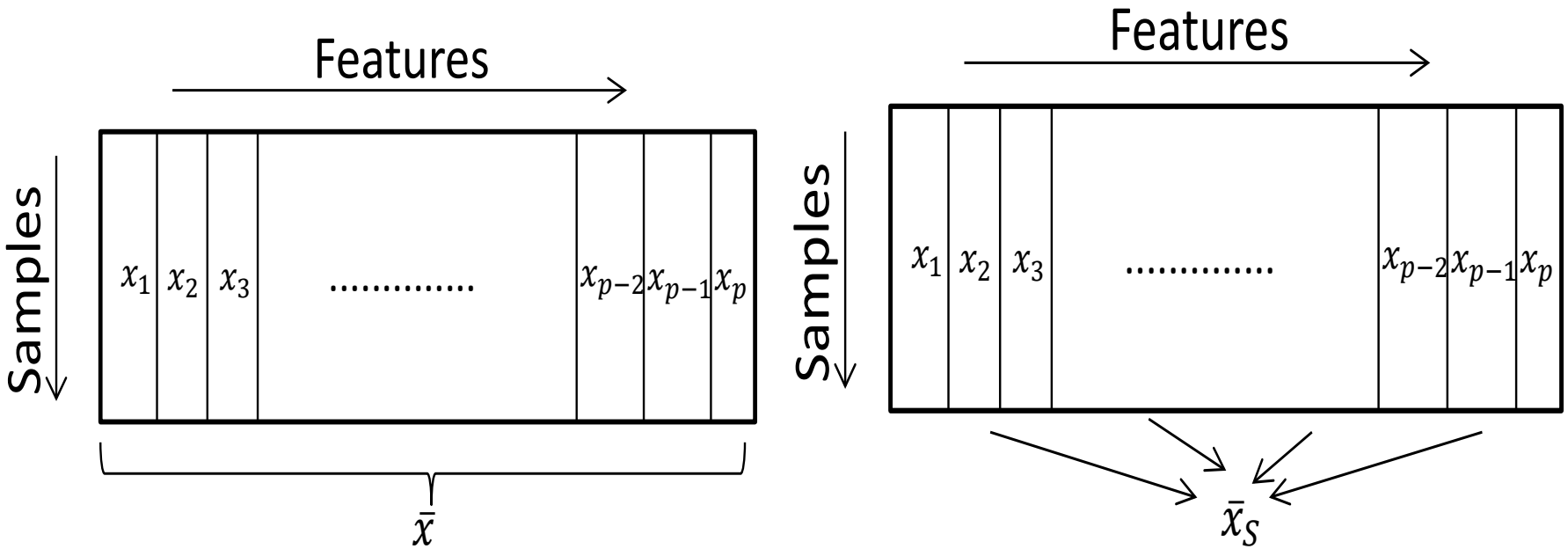


	microarray	two-class model
pre-normalization	yes	**skipped**
feature-wise screening	Kolmogorov-Smirnov	**chi-square**
	$\psi_j = \sup_t \lvert F_{n,j}(t) - \Phi(t)\rvert$	$\psi_j = (\lVert x_j\rVert - n)/\sqrt{2n}$
re-normalization	Efron's null correction	**skipped**
threshold choice	HCT	same
post-selection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

- Notation. $\hat{\ell}_t^{if}$: IF-PCA for a threshold $t > 0$
- Notation. $\hat{\ell}_*^{if}$: classical PCA (a special case)

Aggregation methods

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



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**

*: signals are comparably stronger but still weak

**: a tuning-free version exists

Asymptotic Rare/Weak (ARW) model

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

$$\ell_i = \pm 1 \text{ with equal prob.}, \qquad \mu(j) \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}$$

- "large p small n": $n = p^\theta$, $0 < \theta < 1$
- Rare signals: $\epsilon_p = p^{-\beta}$

Range for sparsity/signal strength: **full** for statistical limits, **more specific** for IF-PCA:

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

Phase function for IF-PCA

Fix $0 < \theta < 1$, define

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

where $\rho(\beta)$ is the standard phase function *[Donoho and Jin (2004), Ingster (1997, 1999), Jin (2009)]*:

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

Phase transition for IF-PCA

Theorem 1. Consider IF-PCA with threshold t. Let $\hat{U}^{(t)} \in \mathbb{R}^n$ be the first left singular vector of post-selection data matrix

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

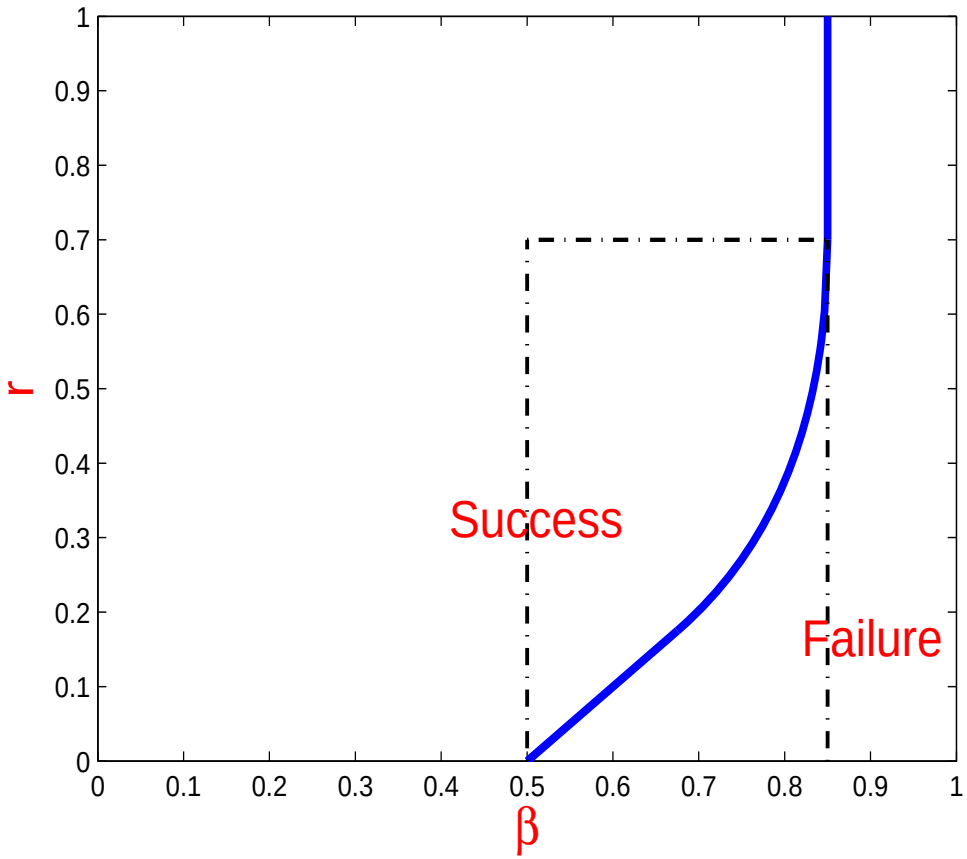
$$\mathrm{Angle}(\hat{U}^{(t)}, \ell) \geq c_0 \qquad \text{(IF-PCA partially fails)}$$

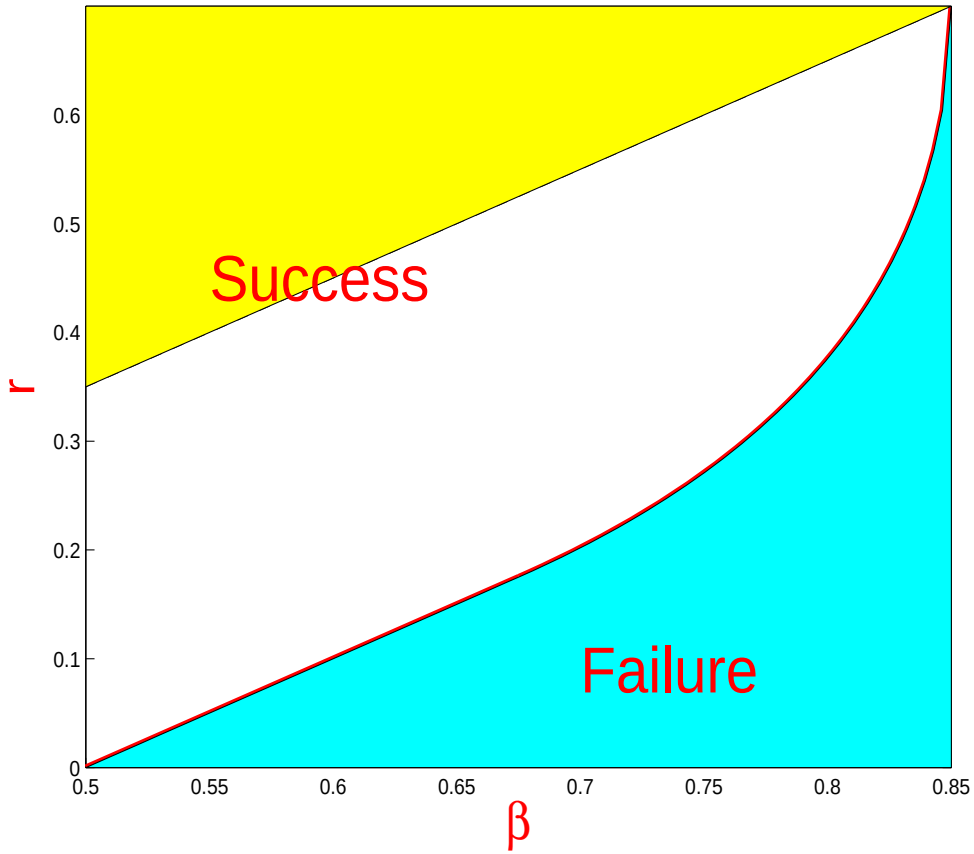
Possibility. If $r > \rho_\theta(\beta)$, then

- For t in an appropriate range, $\mathrm{Angle}(\hat{U}^{(t)}, \ell) \to 0$, and
  $\hat{U}^{(t)} \propto \widetilde{SNR}(t)\ell + z + rem, \qquad \widetilde{SNR}(t) \gg 1, \quad z \sim N(0, I_n)$;
- HCT yields the right threshold choice:
  $t_p^{HC} / t_p^{ideal} \to 1$ in prob., where $t_p^{ideal} = \mathrm{argmax}_t\{\widetilde{SNR}(t)\}$
- IF-PCA-HCT yields successful clustering:
  $$\mathrm{Hamm}_p(\hat{\ell}^{HCT}; \beta, r, \theta) \to 0$$

Phase Diagram (IF-PCA)

$\#(\text{useful features}) \approx p^{1-\beta}$, $\tau_p = \sqrt[4]{(1/n)2r\log(p)}$; $n = p^{\theta}$ $(\theta = .6)$





Statistical limits (clustering)

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

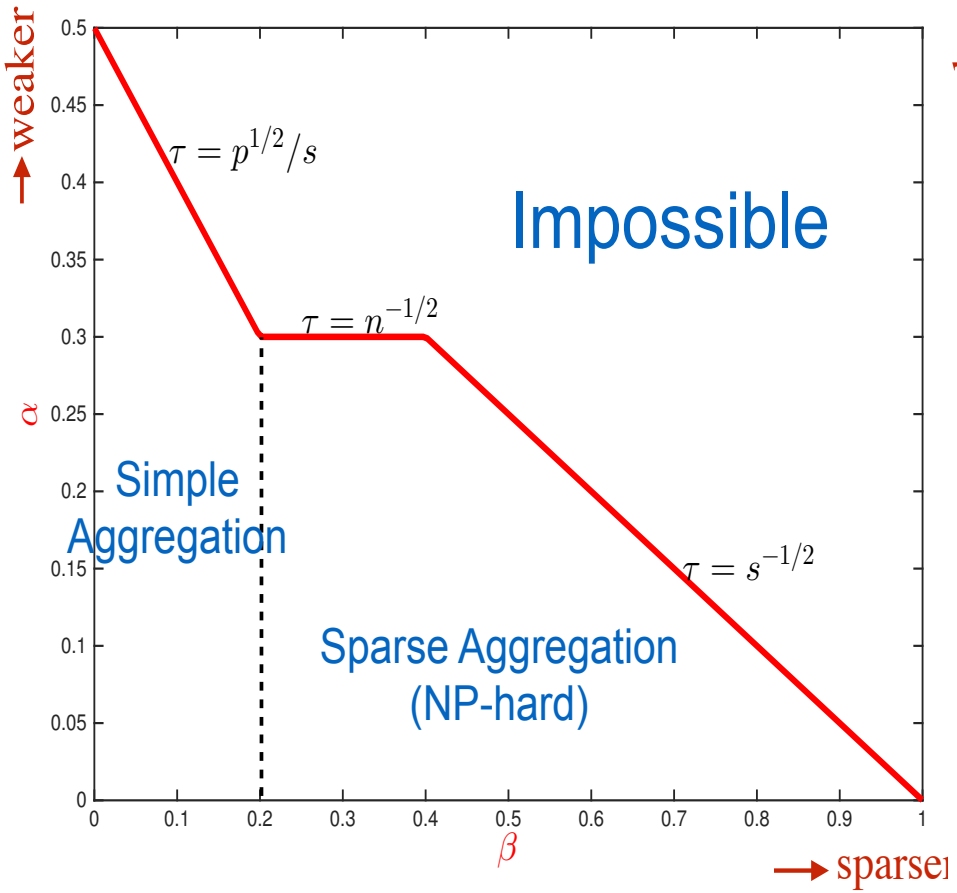
$$\eta_\theta^{clu}(\beta) = \begin{cases} (1-2\beta)/2, & 0 < \beta < \frac{1-\theta}{2} \\ \theta/2, & \frac{1-\theta}{2} < \beta < 1-\theta \\ (1-\beta)/2, & \beta > 1-\theta \end{cases}$$

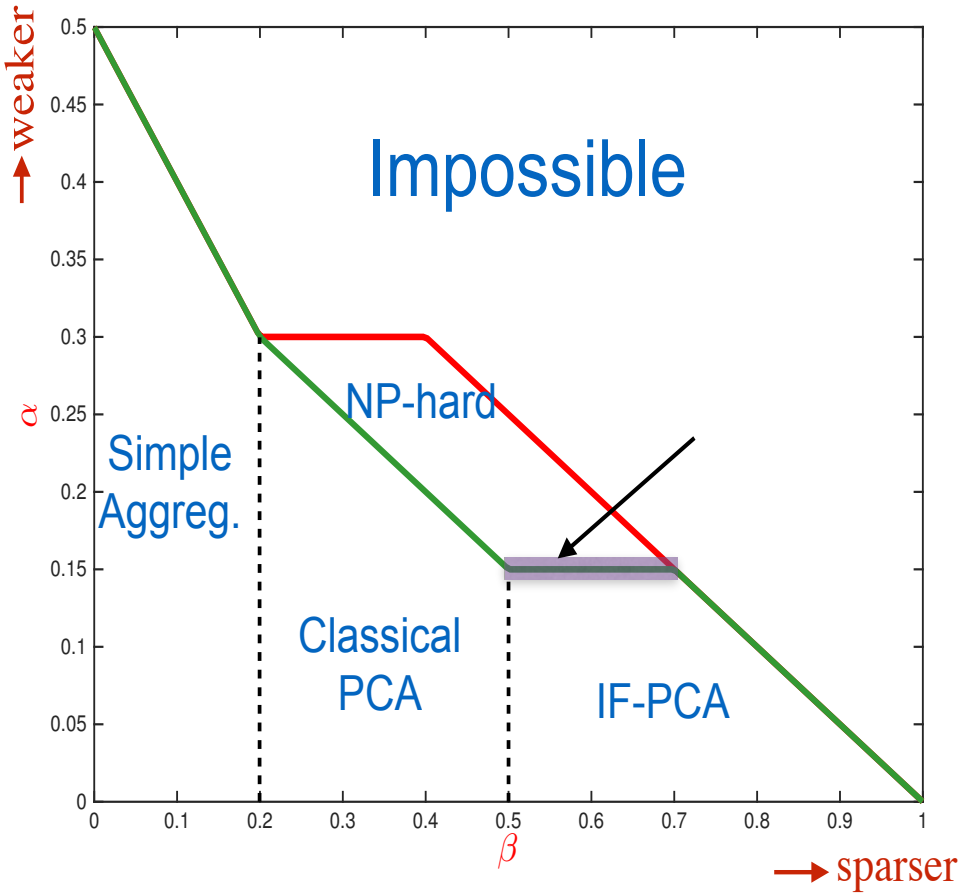
Theorem 1.

- 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 $\beta > \frac{1-\theta}{2}$ and $N = p^{1-\beta}$.

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

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





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)

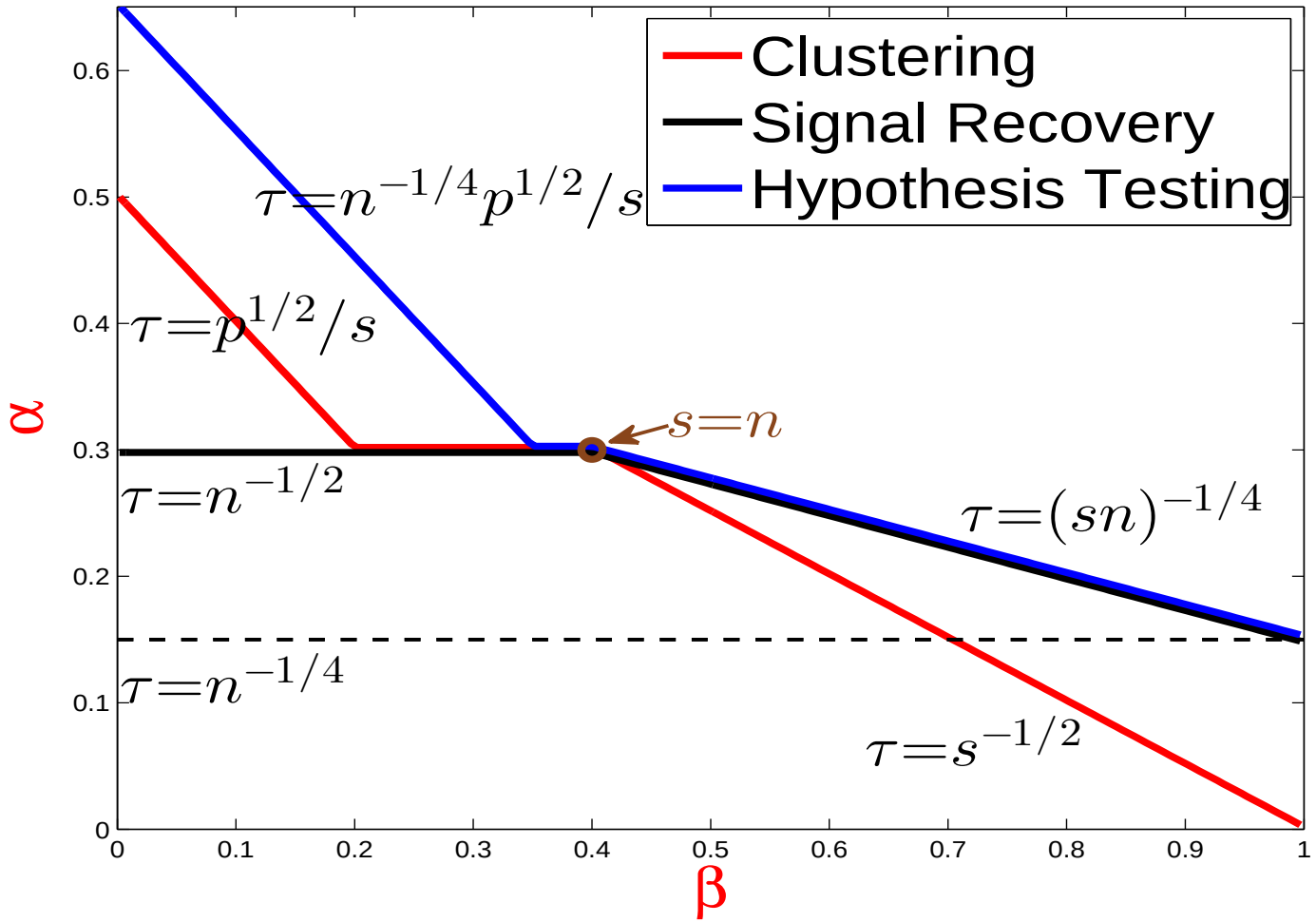
$$\mathit{TestErr}_p(\hat{T}, \beta, r, \beta) = P_{H_0^{(p)}}(\text{Reject } H_0) + P_{H_1^{(p)}}(\text{Accept } H_0^{(p)})$$

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 $= p^{-\alpha}$

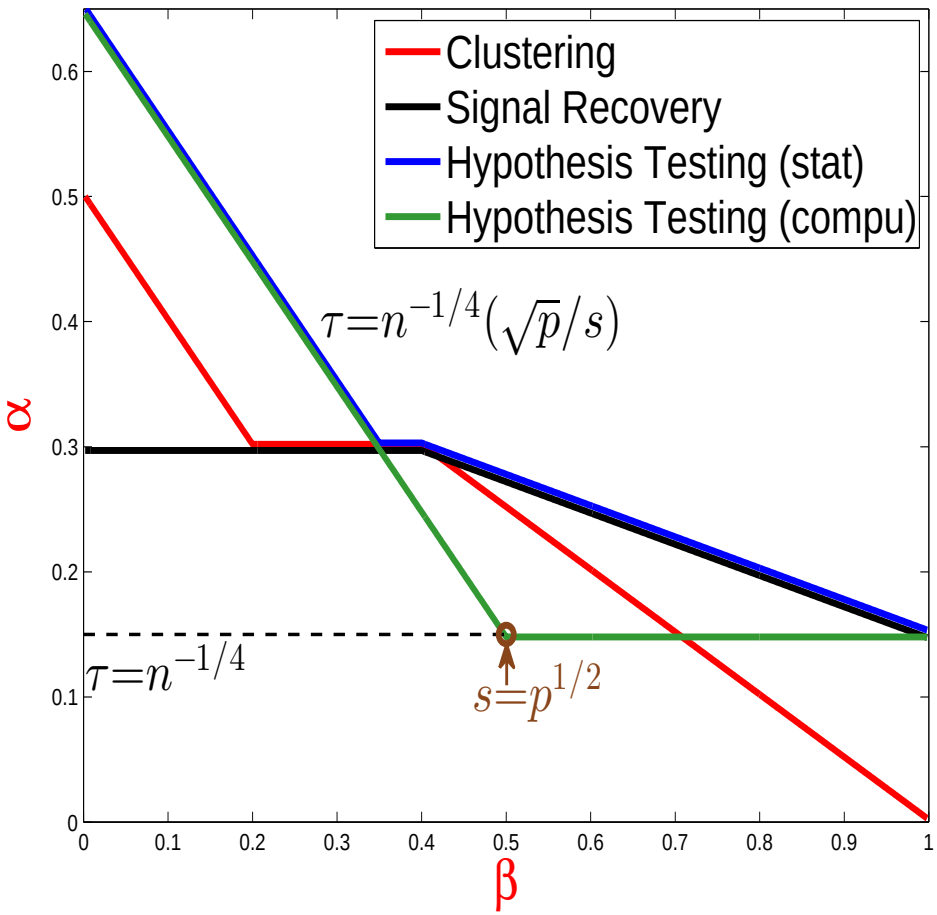


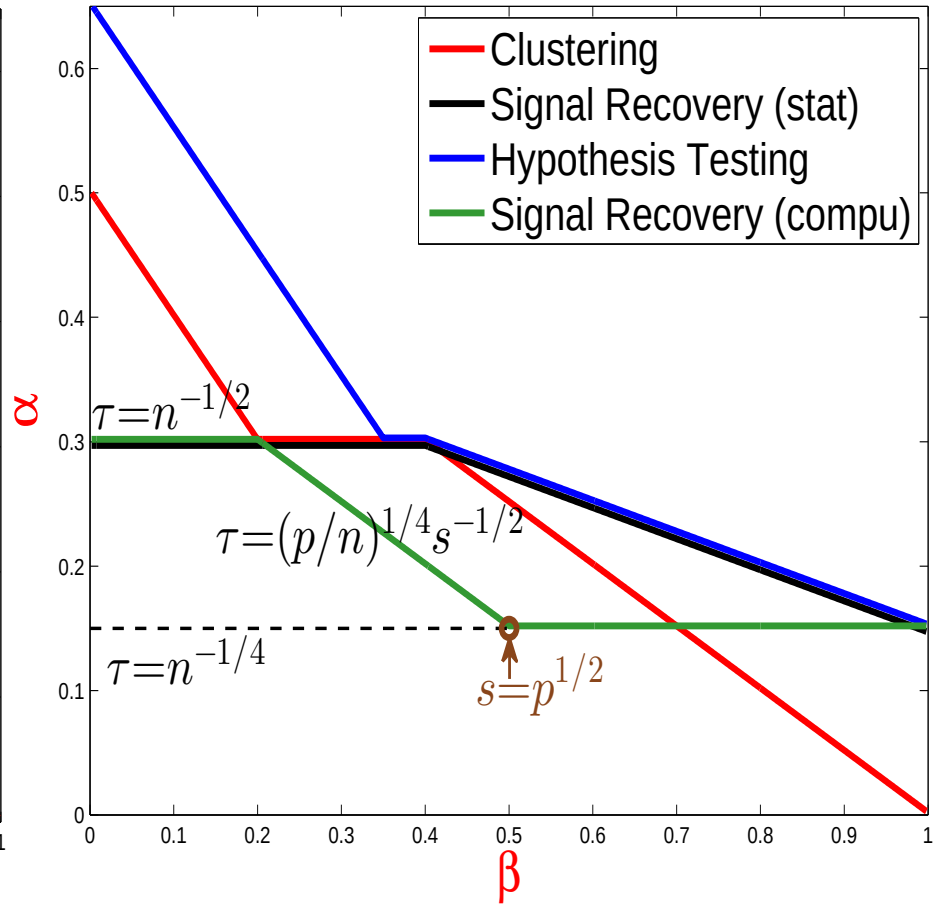
Computable upper bounds (two problems)

Left: feature estimation. 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

www.stat.cmu.edu/~jiashun

- RW settings: found in many scientific problems, need new methods/theory
- Proposed IF-PCA-HCT as an easy-to-adapt and fast method, especially for RW settings
- Showed effective real data results
- Studied phase transitions for IF-PCA, clustering, signal recovery, and global testing

Jin J, Wang W (2014) Important Features PCA for high dimensional clustering (arXiv.1407.5421)
Jin J, Ke Z, Wang W (2015) Phase transitions for high dimensional clustering and related problems (arXiv.1502.06952)