Higher Criticism Thresholding

Optimal Feature Selection when Useful Features are Rare and Weak

David Donoho (s) Jiashun Jin (c)

(c) Carnegie Mellon (s) Stanford

May 17, 2009

Some LDA Background

- 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 feature (X), predict class label Y

Fisher's linear classifier

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

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

Modern Challenges

Iconic examples: gene microarray

Data Name	Source	n, # samples	p,# features
Colon cancer	Alon et al. (99)	62(22, 40)	2000
Leukemia	Golub et al. (99)	73(38,35)	7129
Prostate cancer	Singh et al. (02)	102(50,52)	12600

Problem: Too few observations to estimate Σ^{-1} ($p \gg n$).
Response: use separable classifiers $diag(\Sigma)^{-1}\mu$.

Problem: Many features, most useless, a few useful/weak
Response: feature selection

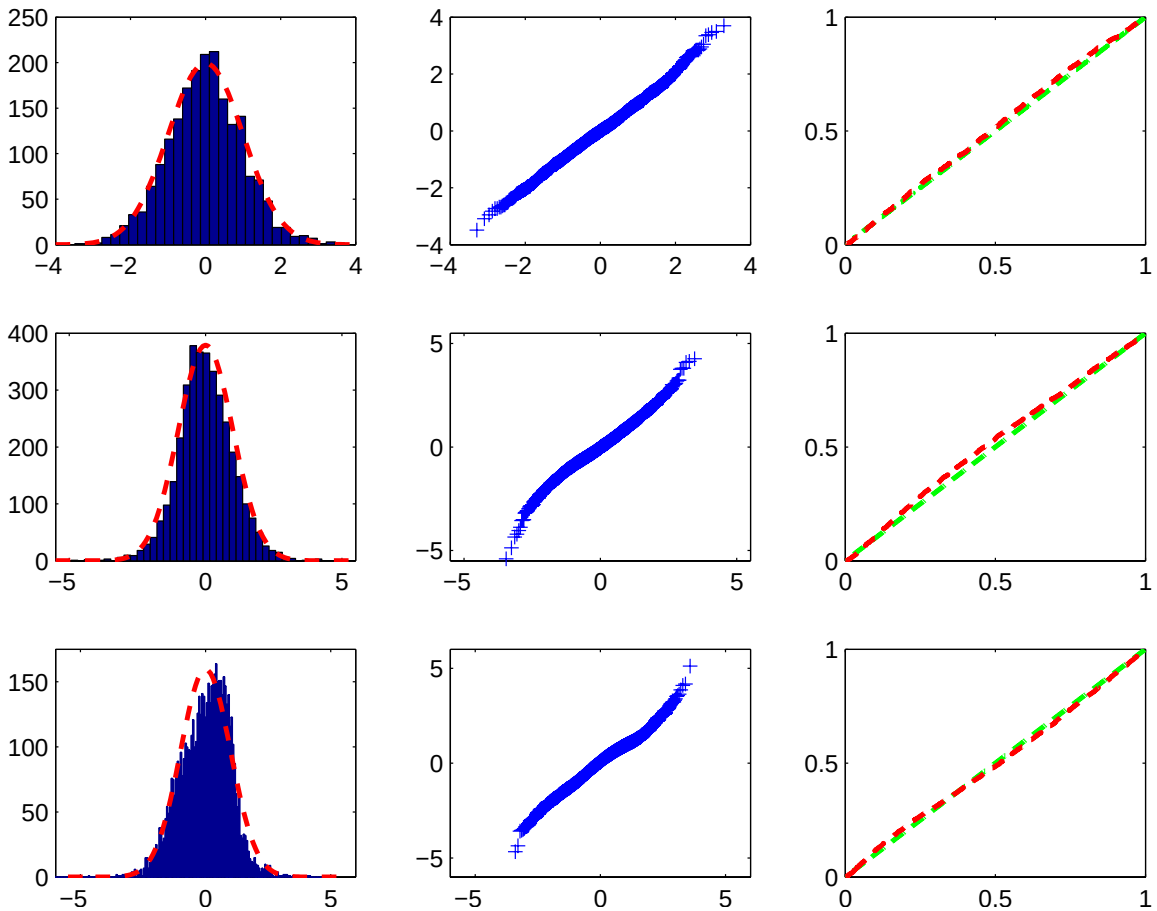
Outcome: Feature Selection + DLDA

e.g. Bickel and Levina (04), Fan &Fan (08), Tibshirani et al. (02)

Feature Selection + DLDA

Step 1. Calculate training Z-vector

- $Z =$ Group Mean Difference$/\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\text{pooled variance}}$
- Standardized by $Z = [Z - mean(Z)]/SD(Z)$



Step 2. Feature Selection by thresholding Z

$$\text{Feature weights:} \quad w_\star^t(j) = \begin{cases} \operatorname{sgn}(Z_j) \cdot 1_{\{|Z_j|>t\}}, & \star = \text{clip} \\ Z_j \cdot 1_{\{|Z_j|>t\}}, & \star = \text{hard} \\ \operatorname{sgn}(Z_j)(|Z_j| - t) \cdot 1_{\{|Z_j|>t\}}, & \star = \text{soft} \end{cases}$$

Step 3. Classification using LDA:

$$L^\star(X;t) = \sum_{j=1}^{p} w_\star^t(j) \cdot \left(\frac{X(j)}{\hat{\sigma}_j}\right) \quad < \quad > \quad 0$$

Problem: What is the best threshold t?

Threshold Choice

Commonly seen intuition:

- low feature FDR (e.g. keep strongest 3 or 5)
- Sure Indep. Screening (SIS) (Fan & Lv 08)
- cross validation (CV)
- threshold monotone with feature strength

For today:

- Threshold choice by Higher Criticism (**HC**)
- Re-investigate the above ideas

Outline

- Higher Criticism Thresholding (HCT)
- Insight, and Rare Weak Model (RW)
- Phase diagram/Optimality (Asymptotic RW)
- Comparison with FDRT/SIS/CVT

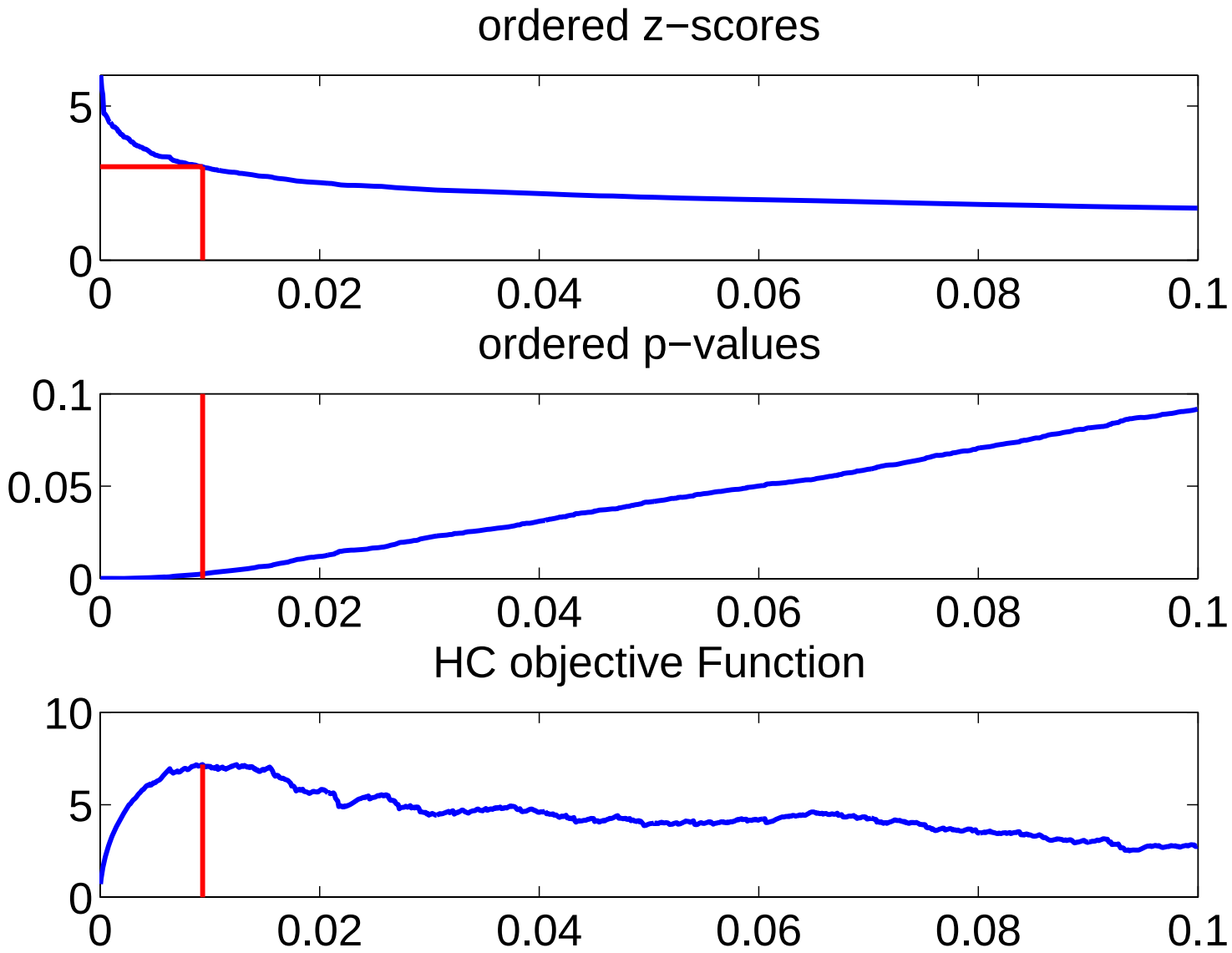
Higher Criticism Threshold (HCT)

Z_j : z-score for testing whether j-th feature is useful

1. Convert to P-values: $\pi_j = P\{|N(0,1)| > |Z_j|\}$
2. Sort: $\pi_{(1)} < \pi_{(2)} \ldots < \pi_{(p)}$
3. HC objective function $HC^*_{n,p} = \max_{1 \le i \le \alpha_0 \cdot p} \Big\{ \sqrt{p} \Big(\frac{\frac{i}{p} - \pi_{(i)}}{\sqrt{i/p(1-i/p)}} \Big) \Big\}$
4. HC-threshold (HCT): **(new ingredient)**

$$t_{HC} = |Z|_{(\hat{i})} \text{ corresponding to maximizing } i$$

Note: (1). slight difference of HC from Donoho & Jin 04. (2). Hall et al 08 uses HC for classification without features selection; see Donoho & Jin 08 for comparison


ordered z−scores
5
0
0
0.02
0.04
0.06
0.08
0.1
ordered p−values
0.1
0.05
0
0
0.02
0.04
0.06
0.08
0.1
HC objective Function
10
5
0
0
0.02
0.04
0.06
0.08
0.1

Comparison with Popular Classifiers

Data: Leukemia/Colon/Prostate

- $(2/3, 1/3)$ random split (Train, Test).
- average test errors across 50 replications
- $\text{regret} = \dfrac{\text{Cell value} - \text{Column min}}{\text{Column max} - \text{Column min}}$

All except that of HC is from Dettling's paper.

Method	Colon	regret	Leukemia	regret	Prostate	regret	Max. Regret	Rank
Bagboost	16.10	.58	4.08	.59	7.53	0	.59	4
Boosting	19.14	1	5.67	1	8.71	.13	1.00	7.5
RanFor	14.86	.41	1.92	.02	9.00	.41	.41	2
SVM	15.05	.44	1.83	0	7.88	.04	.44	3
PAM *	11.90	0	3.75	.50	16.54	1	1.00	7.5
DLDA	12.86	.13	2.92	.28	14.18	.74	.74	6
KNN	16.38	.62	3.83	.52	10.59	.34	.62	5
HCT-hard	13.77	.26	3.02	.31	9.47	.22	.31	1

* Tibshirani et al. posted very different figures.

See Donoho and Jin (2008) for comparison with simulated results

Rare/Weak Features Model (RW)

- n training samples (X_i, Y_i):
  $X_i \sim N(Y_i \cdot \mu,\ \Sigma)$, $Y_i = \pm 1$: class lables
- Z-vector: $Z \sim N(\sqrt{n} \cdot \mu,\ \Sigma)$
- test feature: $X \sim N(\pm\mu,\ \Sigma)$

RW model:

- $\Sigma = I_p$
- $\sqrt{n} \cdot \mu_j = \begin{cases} \tau, & j\text{-th feature is useful} \\ 0, & j\text{-th feature is useless} \end{cases}$
- $\epsilon = \frac{1}{p} \cdot \#\{j :\ \mu_j \neq 0\}$

Four key parameters:

$p \gg n$, $\epsilon \approx 0$, τ small or moderately large

Idea

Definition

- **Optimal** threshold: minimizes $P\{\text{misclassified} \mid t\}$
- **Ideal** threshold: minimizes a proxy of $P\{\text{misclassified} \mid t\}$
- **HCT**: maximizes HC objective function
- **Ideal HCT**: maximizes Ideal HC objective function

Key: in a broad situation (including RW Model)

Optima threshold $\approx$ Ideal threshold $\approx$ Ideal HCT $\approx$ HCT

Insight I, Fisher's Separation

Linear Classifier score $L(X) = w'X$.

$$SEP(L;\mu) = \frac{(\text{Diff. of mean scores} \mid \mu)}{\sqrt{(\text{Variance of scores} \mid \mu)}} = \frac{w'\mu}{\|w\|_2}$$

- Clip: $L_t(X) = \sum \mathrm{sgn}(Z_j) \cdot 1_{\{|Z_j| \geq t\}} \cdot X(j) \quad < \ > 0$
- $P\{\text{misclassified } |t\} = E_{\epsilon,\tau} E_Z[\bar{\Phi}(SEP(L_t \mid \mu))]$
- **IF** order of "E" and "$\bar{\Phi}$" can be interchanged:

$$E_{\epsilon,\tau} E_Z[\bar{\Phi}(SEP(L_t;\mu))] \approx \bar{\Phi}(\widetilde{SEP}(t))$$

where $\widetilde{SEP}(t) = (EL_t(\mu))/||EVar(L_t(X)|\mu)||_2$

THEN Optimal threshold $\approx$ Ideal threshold

Signal Detection Background

Positives: call a training z-score Z_i a positive if

$$|Z_i| \geq t$$

Positive Rate (PR):

$$PR(t) \equiv 2(1-\epsilon)\bar{\Phi}(t) + \epsilon\bar{\Phi}(t-\tau) + \epsilon\bar{\Phi}(t+\tau)$$

True Positive Rate (TPR)

$$TPR(t) = \epsilon \cdot [\bar{\Phi}(t-\tau) + \epsilon\bar{\Phi}(t+\tau)]$$

note: both are expected values

Insight II, Intimacy of SEP and HC

- Neglect stochastic fluctuations, HC reduces to Ideal HC:

$$\widetilde{HC}(t;\epsilon,\tau) = \frac{\epsilon \cdot [\bar{\Phi}(t-\tau) + \bar{\Phi}(t+\tau) - 2\bar{\Phi}(t)]}{\sqrt{PR(t)(1-PR(t))}}$$

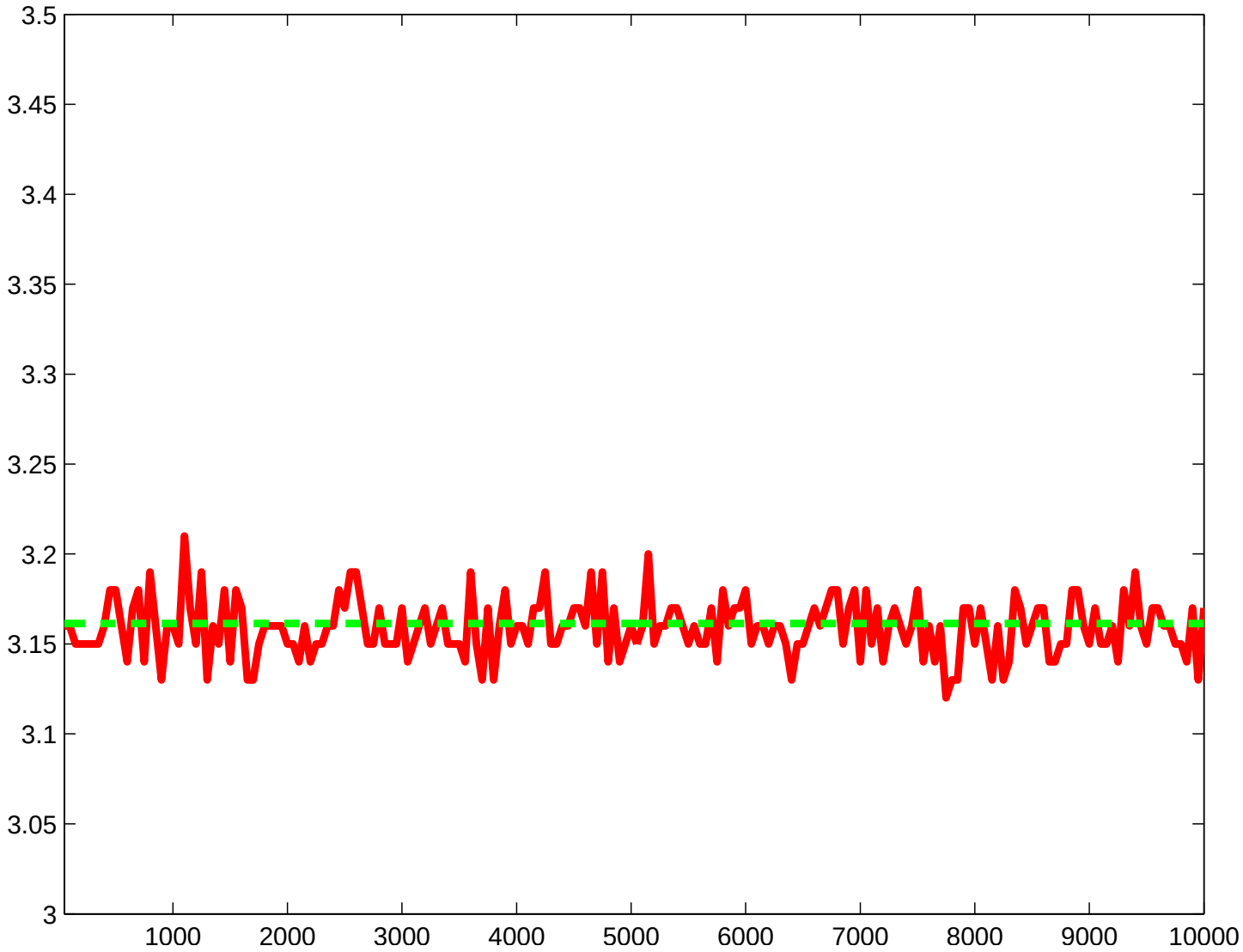
- Ideal Thresholding: maximize

$$\widetilde{Sep}(t;\epsilon,\tau) = \frac{\epsilon \cdot [\bar{\Phi}(t-\tau) - \bar{\Phi}(t+\tau)]}{\sqrt{PR(t)}} \approx \frac{\epsilon \cdot TPR(t)}{\sqrt{PR(t)}}$$

- In RW Model, parameters $\epsilon \approx 0$, τ moderate to large, so

$$\widetilde{HC}(t;\epsilon,\tau) \approx \widetilde{Sep}(t;\epsilon,\tau) \approx \frac{\epsilon \cdot TPR(t)}{\sqrt{PR(t)}}$$

- Optimal threshold $\approx$ Ideal threshold $\approx$ Ideal HCT $\approx$ HCT



Green: Average HCT over 100 simulations; Red: Optimal threshold

$p = 10,000$, $\epsilon = 0.01$, $\tau = 3.5$, n ranges from 50 to $10,000$

Asymptotic Rare/Weak Model (ARW)

Number of features p grows to ∞

- Linking rarity/weakness to p:

$$\epsilon_p = p^{-\beta}, \qquad 0 < \beta < 1$$

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

- Linking 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 $\theta \in (0, 1)$

Impossibility and Possibility

Introduce

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

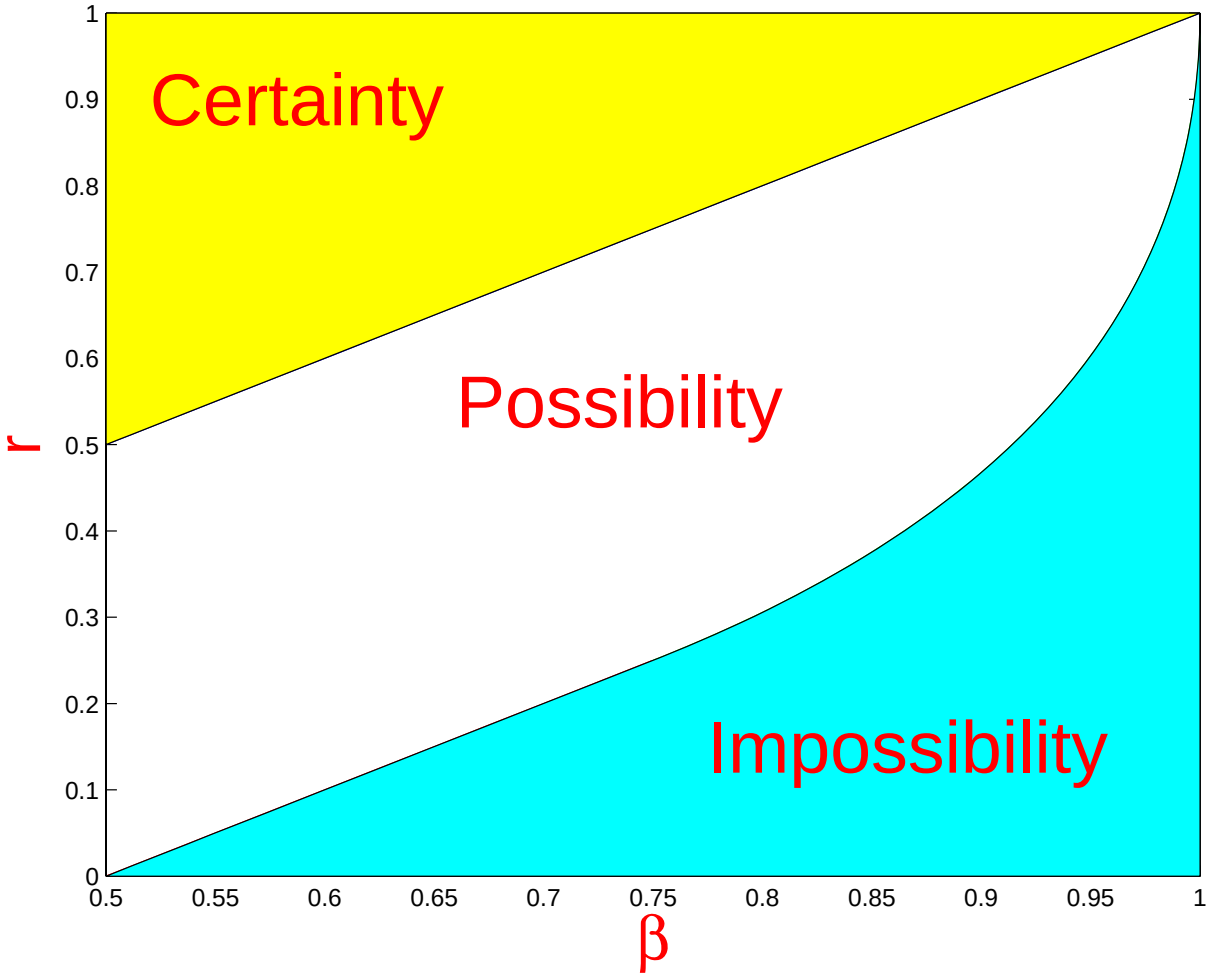
and

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

$r = \rho^\star(\beta)$ partitions β-r plane into two regions:

Region of Possibility, Region of Impossibility

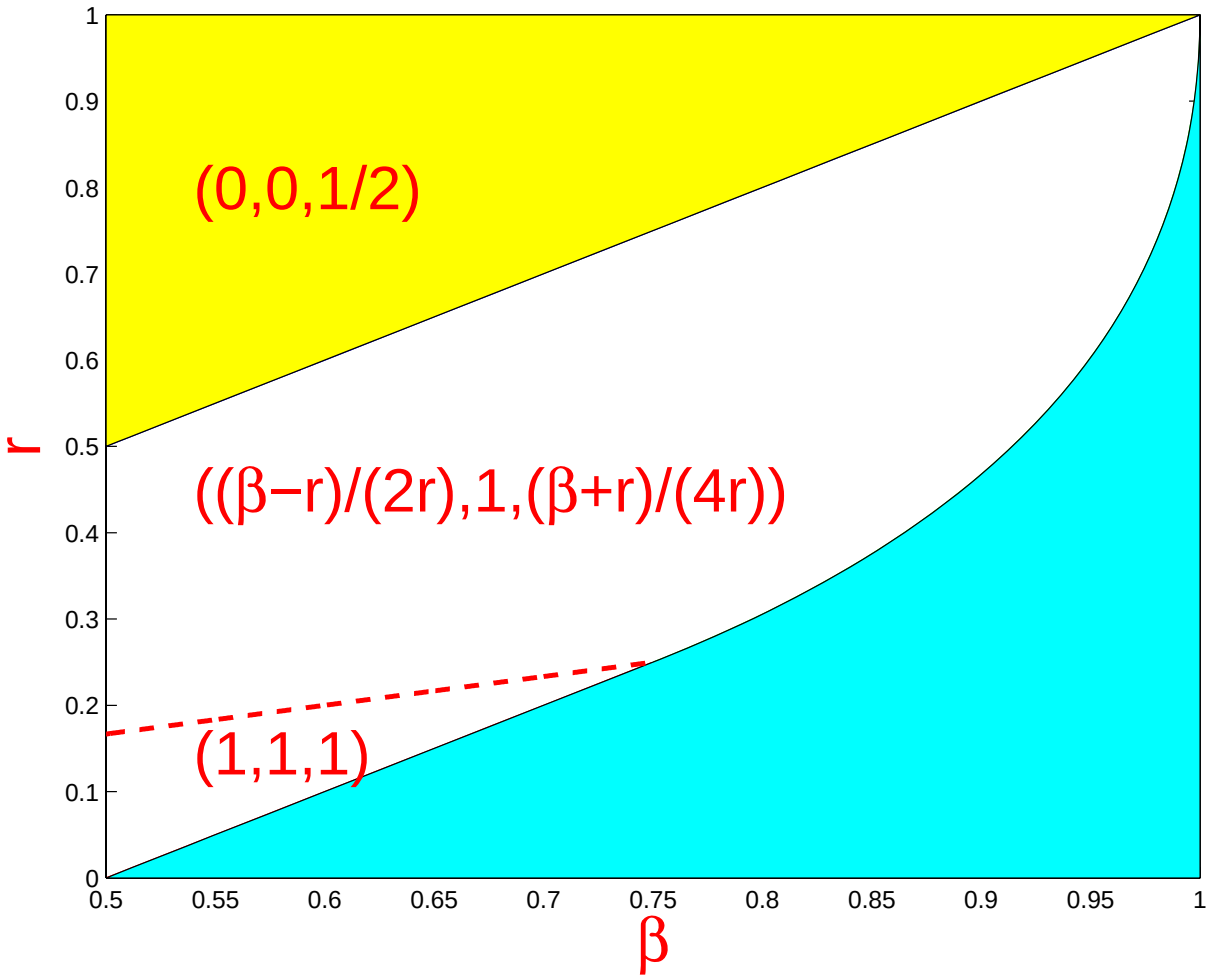
Phase Diagram (Slow Growth)



$$\epsilon = p^{-\beta}, \qquad \tau = \sqrt{2r \log p}, \qquad 1/2 < \beta < 1, \qquad 0 < r < 1$$

Region of Impossibility: help to explain failure of reproducibility

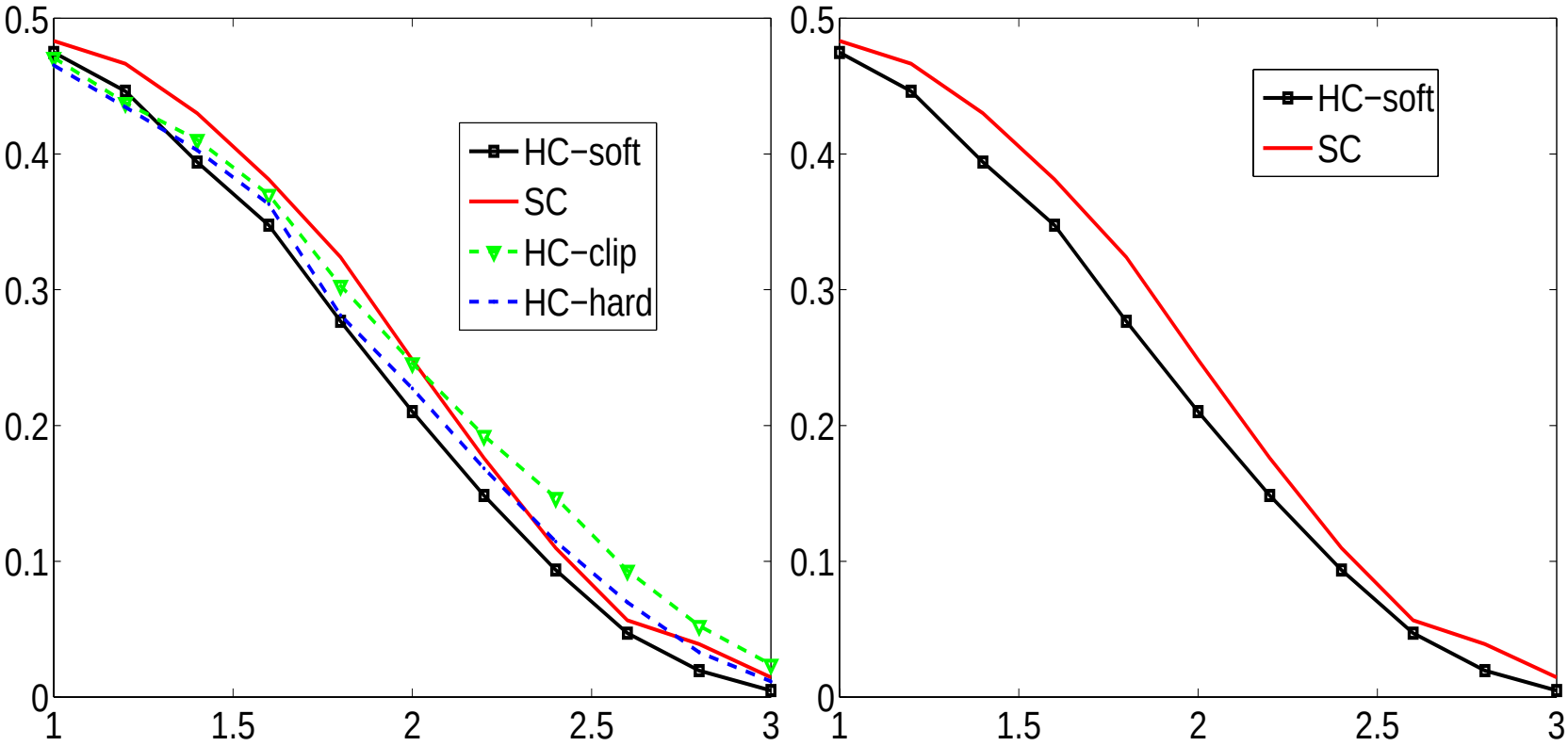
Comparison: HCT vs. FDRT and SIS



$$\epsilon = p^{-\beta}, \qquad \tau = \sqrt{2r \log p}, \qquad 1/2 < \beta < 1, \qquad 0 < r < 1$$

Number in brackets: (FDR, MDR, local FDR)

Comparison to Shrunken Centroid (CVT)



$p = 10^4$, $n = 40$;

100 useful features generated from $N(\tau/\sqrt{n}, 1)$, $\tau \in [1, 3]$;

9900 useless features generated from $N(0, 1)$

Take-home messages

- New threshold for feature selection when useful features are rare and weak (RW) in the large-p, small-n setting
- Optimal classification performance
- Very different from fashionable FDRT
- Can replaced CVT with lower cost and better performance
- Competitive on standard real datasets

Acknowledgement: We thank Issac Newton Institute for hospitality

Website

www.stat.cmu.edu/~jiashun/Research/

Available: DLD & JJ (2008): definition, heuristics, practical results
JJ (2009): region of possibility/impossibility
DLD & JJ (2009): phase diagram, first order asymptotics

In preparation: full achievability, extensions, second order asymptotics