Higher Criticism Thresholding

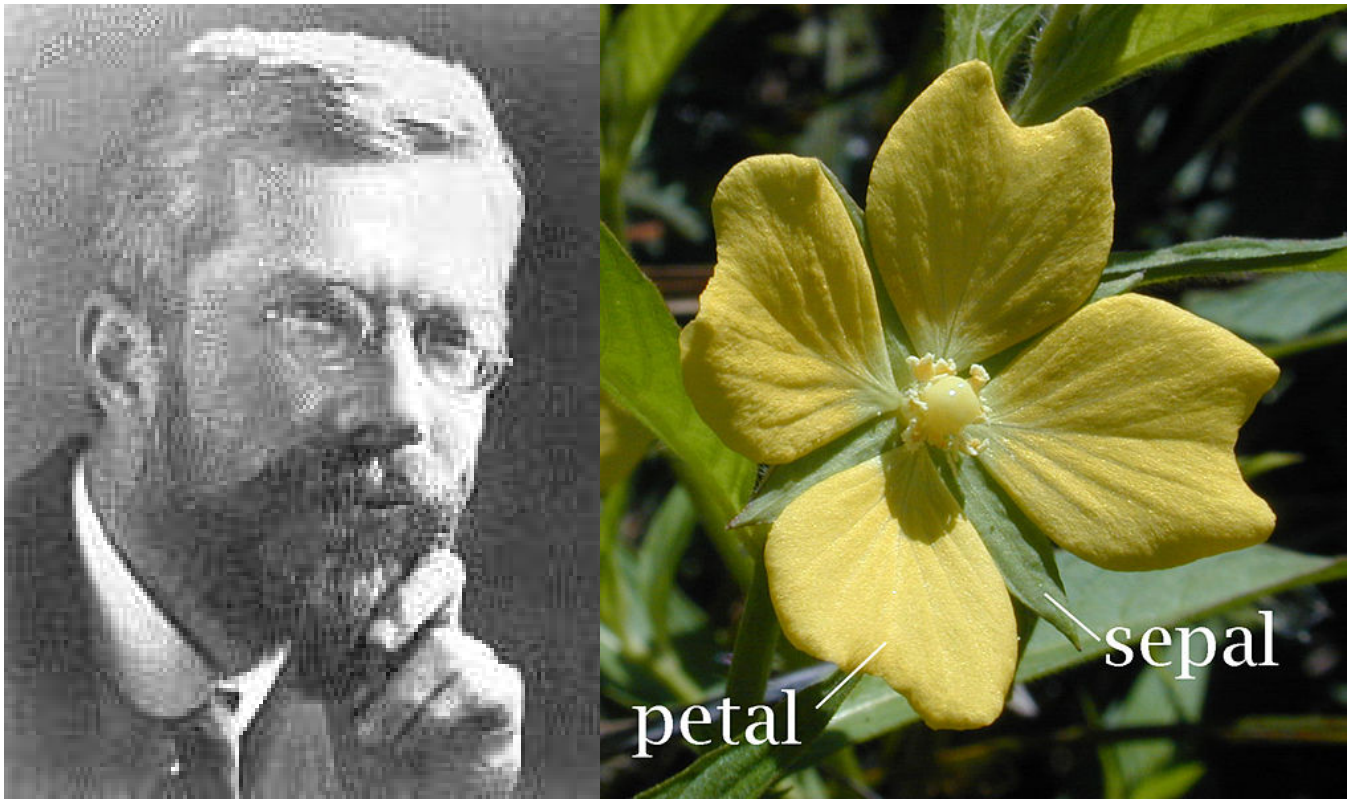
Optimal Feature Selection when Useful Features are Rare and Weak

Jiashun Jin

Carnegie Mellon University

With David Donoho (Stanford)

Fisher's Iris Data



Sir Ronald Fisher (1890-1962); Iris flower (50 samples, 4 features)

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 $Y = 1$ or $Y = -1$

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

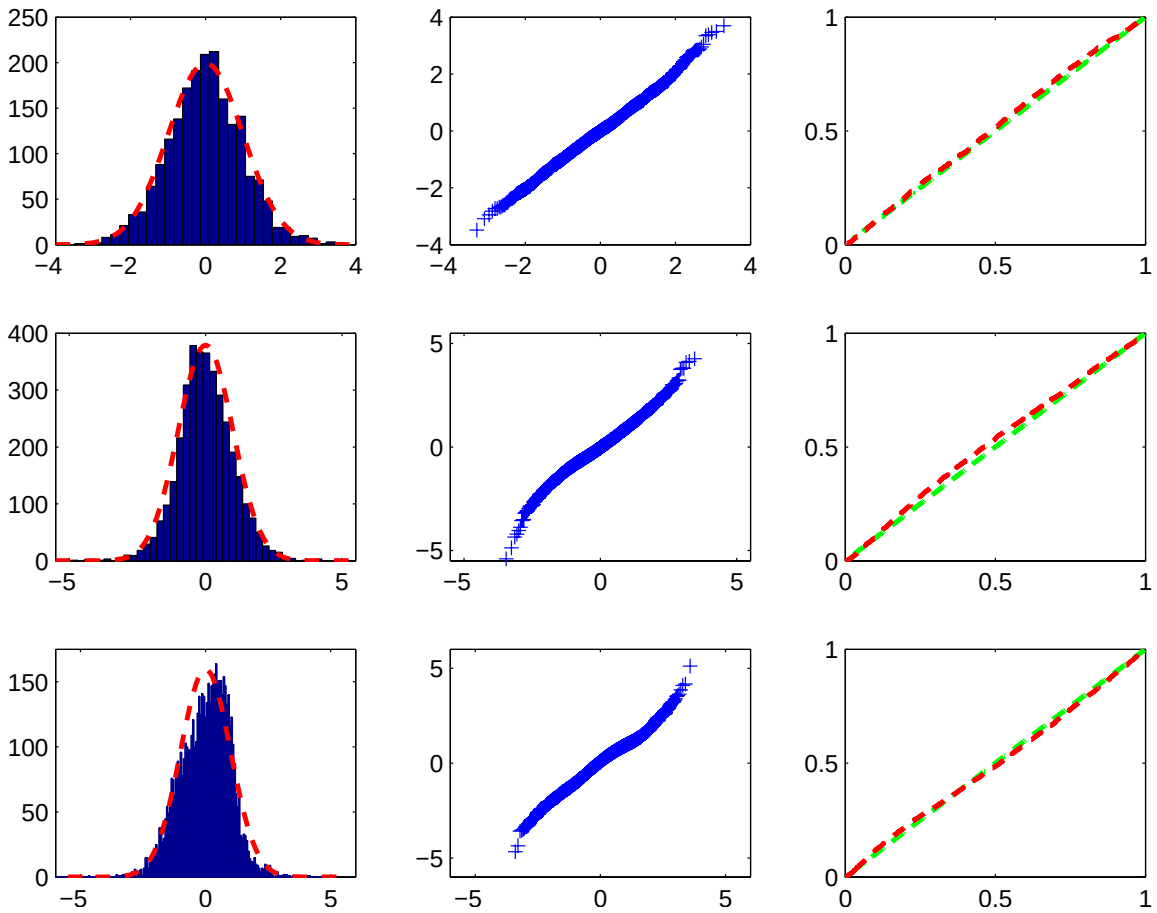
Outcome: Feature Selection + DLDA

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

DLDA with Feature Selection

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

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

Step 3. Classification using LDA:

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

Problem: What is the best threshold t?

Threshold Choice

Commonly seen intuition:

- cross validation (CV)
- control feature FDR (many)
- control feature Lfdr (Efron & others)
- Sure Indep. Screening (SIS) (Fan & Lv 2008)
- threshold monotone with feature strength

For today:

- threshold choice by Higher Criticism (**HC**)
- optimality of this choice
- re-investigate the above ideas

Higher Criticism (HC)

- a multiple testing notion by Tukey (1976)
- optimal in detecting very sparse Gaussian mixture (Donoho & Jin 2004)
- useful in Cosmology/Genomics/Comp.Sensing
  - Jin et al. (2005); Cayon et al. (2006)
  - Goeman & Bulhmann (2007)
  - Nowak et al. (2009)
- extended to many cases
  - Meinshausen & Rice (2006); Cai et al. (2007)
  - Jager & Wellner (2007)
  - Hall & Jin (2008, 2009); Delaigle & Hall (2008)
- related ideas: Kendall & Kendall (1982), S. Holm (1982)

Tukey's Stat411 Notes 1976

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 --

Tukey's Story (In Context of Multiple Testing)

- Example: A young scientist administers 250 uncorrelated tests, out of which 11 were significant at the 5% level.
- Question is: Is this surprising?
- Answer: No, we expect

$$250 \times 5\% = 12.5$$

significance at 5% level

Higher Criticism, Formalization by Tukey

- Higher Criticism statistics:

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

and typically,

$$\text{Reject } H_0 \text{ if and only if } \quad HC_{.05,p} \geq 2$$

- Solution to previous example:

$$\text{Accept } H_0: \qquad HC_{.05,p} = \sqrt{250}\frac{(\frac{11}{250}) - 0.05}{\sqrt{.05 \times .95}} = -.43$$

Higher Criticism Threshold (HCT)

DJ (2008, PNAS)

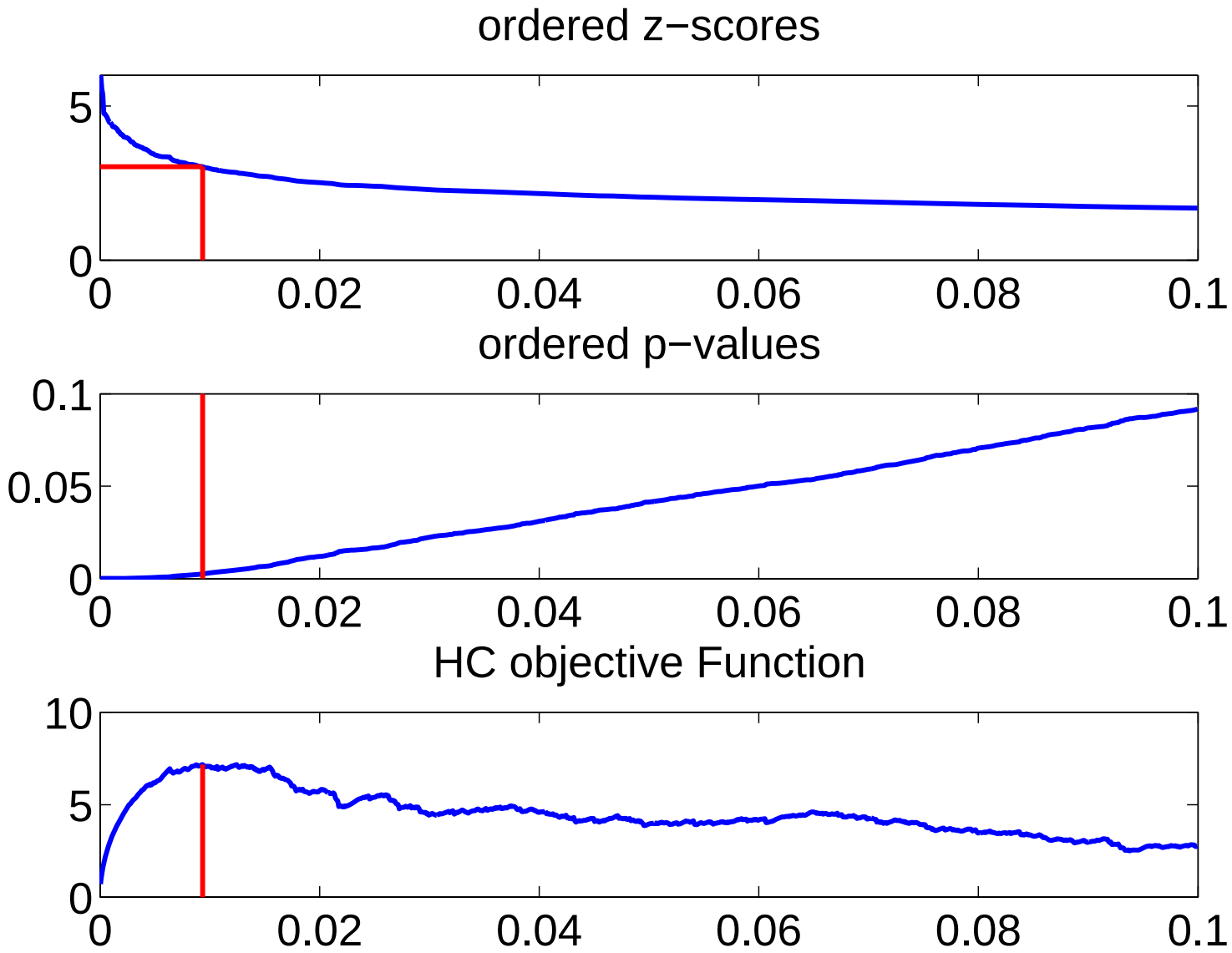
Z_j : z-score for testing whether j-th feature is useful

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

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

Note: (1). slightly different from the HC in DJ (2004, AoS)

(2). Hall et al 08 uses HC for classification without features selection


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.

Comparison with Popular Classifiers, II

- Datasets ...
  - standard in methodological lit. Dettling (2004).
  - not selected by/for us
- HCT ...
  - extremely simple and extremely fast.
  - competitive in misclassification rate
- Other methods
  - Require tuning,
  - Require cross-validation,
  - Internally very complex,
  - .... but don't outperform.

Comparison with simulated data: see DJ (2008, PNAS)

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:

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

4 key parameters:

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

Idea

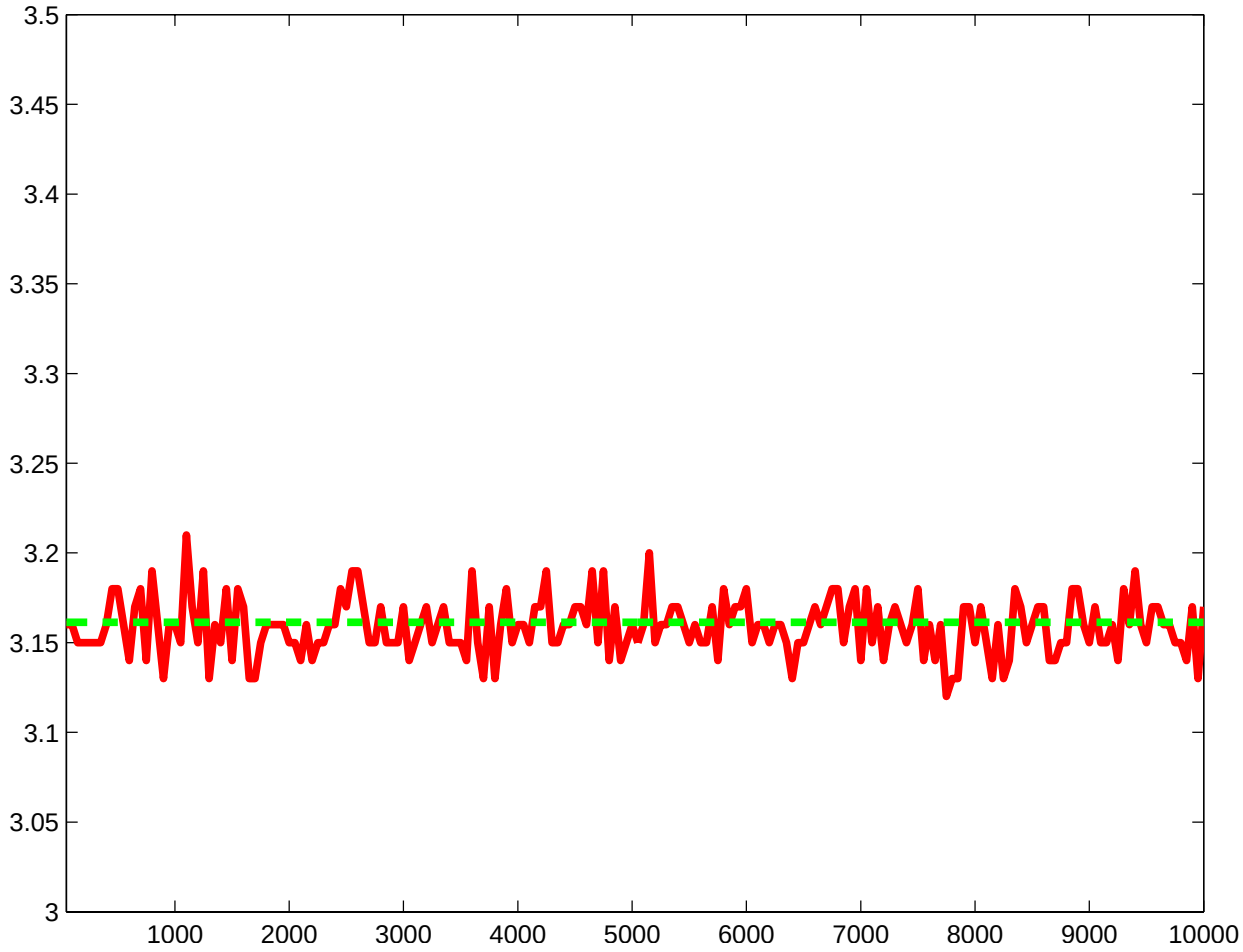
DJ (2009, Phil. Trans. Roy. Soc. A)

Definition

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

Key: in a broad situation (including RW Model)

$$\text{Optima threshold} \approx \text{Ideal threshold} \approx \text{Ideal HCT} \approx \text{HCT}$$



x-axis: n ranges from 50 to $10,000$

y-axis: Average HCT over 100 runs (green) and Optimal threshold (red)

$p = 10,000$ features, 100 are useful, each with strength $\tau = 3.5/\sqrt{n}$

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

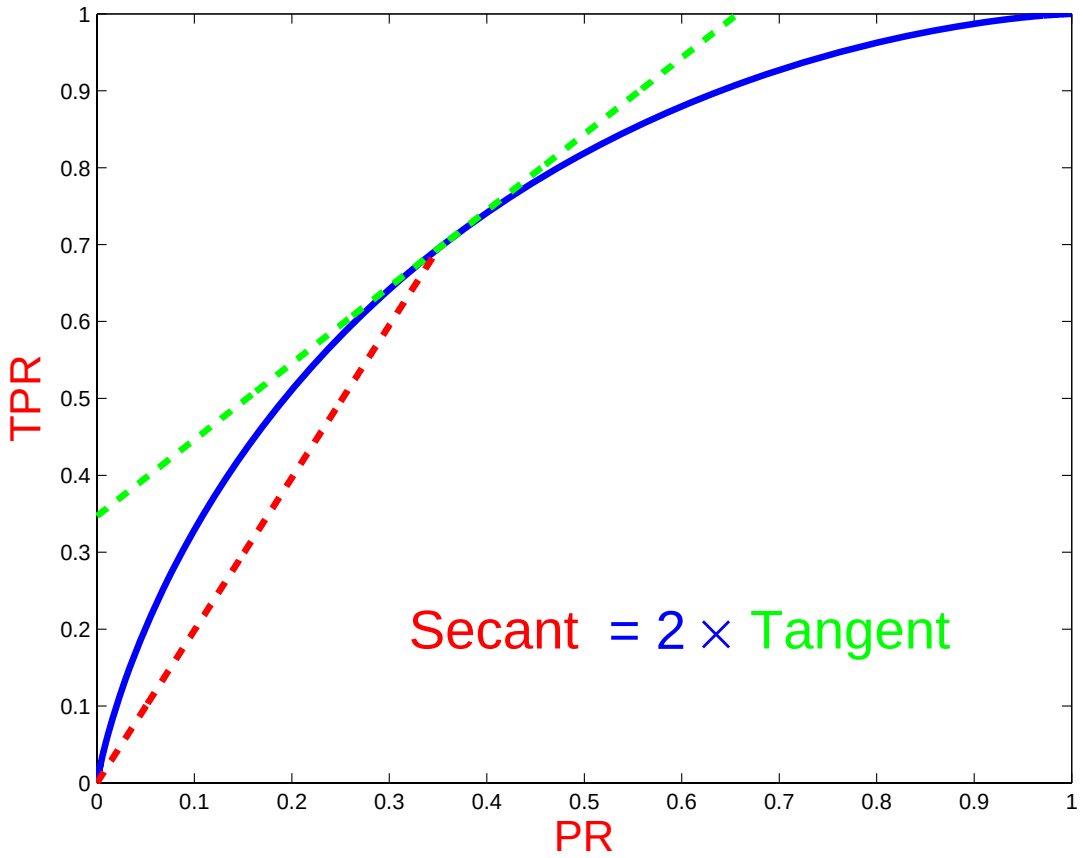
$$\widetilde{HC}(t;\epsilon,\tau) \approx \widetilde{Sep}(t;\epsilon,\tau) \approx \frac{\epsilon\cdot TPR(t)}{\sqrt{PR(t)}}$$

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

A Surprising Connection to "ROC"

$$\widetilde{HC}(t) \approx \widetilde{SEP}(t) \approx \frac{\epsilon \cdot TPR(t)}{\sqrt{PR(t)}}$$

Taking derivative at maximizing value t: $\dfrac{TPR(t)}{PR(t)} = 2 \times \dfrac{TPR'(t)}{PR'(t)}$ (twice rule)



Asymptotics

Recall RW model:

- training-Z-vector: $Z \sim N(\sqrt{n} \cdot \mu,\ I_p)$
- test feature: $X \sim N(\pm\mu,\ I_p)$
- $\sqrt{n} \cdot \mu_j = \begin{cases} \tau, & j\text{-th feature is useful} \\ 0, & j\text{-th feature is useless} \end{cases}$
- Four key parameters

$$p \gg n, \qquad \epsilon \approx 0, \qquad \tau \text{ small or moderately large}$$

For asymptotic study,

- we link ϵ, τ, n to p through some parameters
- then let $p \to \infty$

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 $\theta \in (0, 1)$

Sparse Classification Boundary

J (2009, PNAS)

Define

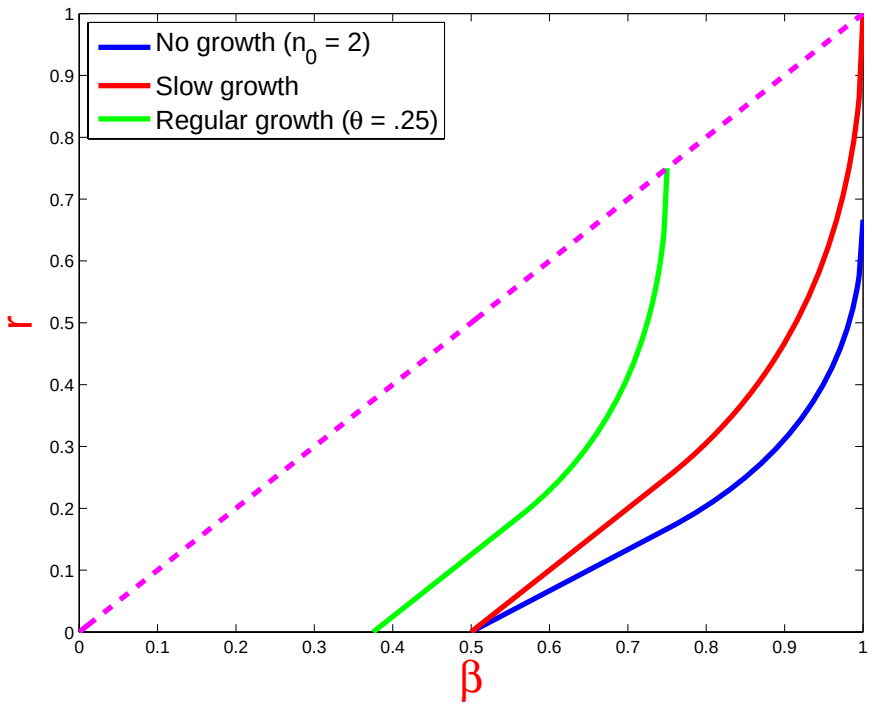
$$\rho(\beta) = \begin{cases} 0, & 0 < \beta < 1/2 \\ (\beta - 1/2), & 1/2 \le \beta < 3/4 \\ (1 - \sqrt{1-\beta})^2, & 3/4 \le \beta < 1 \end{cases}$$

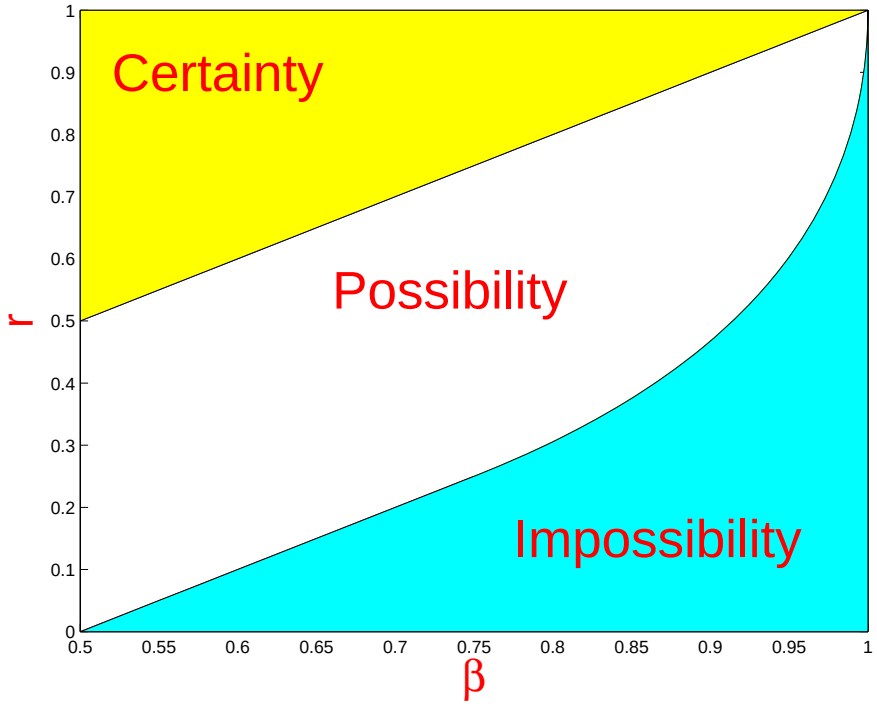
Let

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

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

Phase Diagram





Left: Classification Boundaries. Right: Phase diagram (slow growth)

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

Region of Impossibility: All Classifiers Fail

- Fix a growth type and fix a point (β, r) in the corresponding Region of Impossibility
- Consider a sequence of problems $ARW(r, \beta, n_p)$
- Consider a sequence of classifier training methods (perhaps also dependent on p)

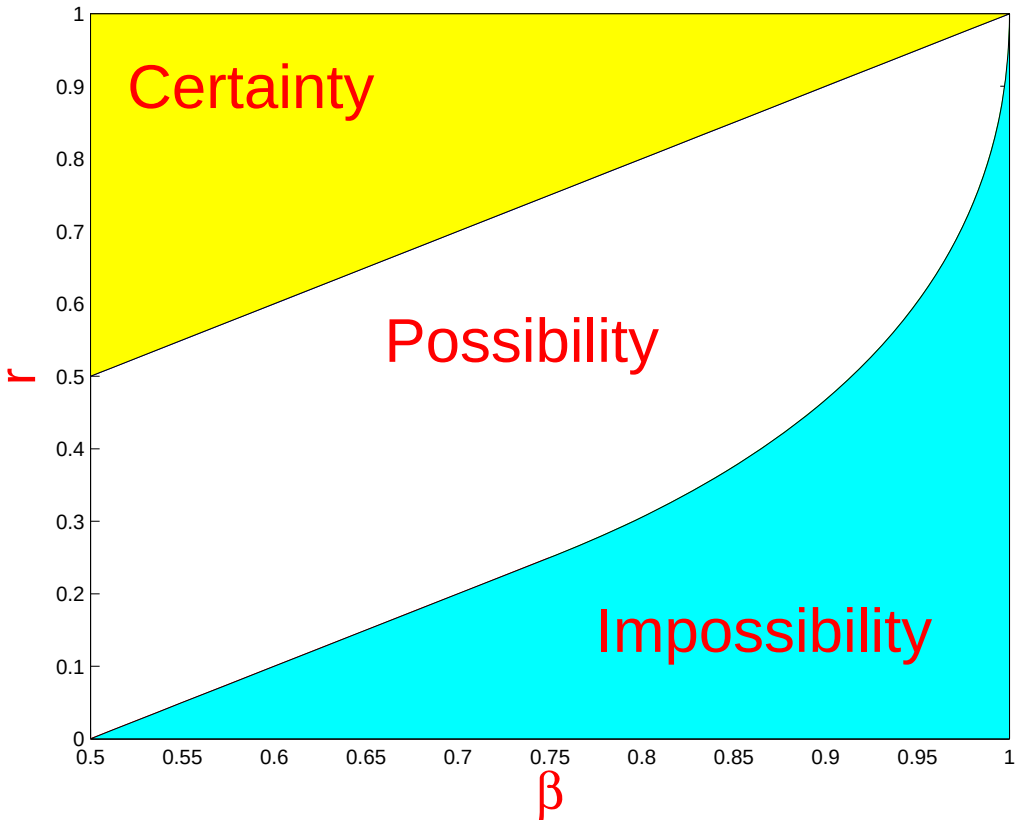
Theorem 1. *The misclassification error rate of the resulting sequence of trained classifiers* $\to 1/2$ *with increasing* p.

Region of Possibility: HCT Classify Successfully

- Fix a growth type and fix a point (β, r) in the corresponding Region of Possibility
- Fix a sequence of problems $ARW(r, \beta, n_p)$
- Consider each of the three training classifiers HCT-clip, HCT-soft, and HCT-hard

Theorem 2. *The misclassification error rate* $\to 0$ *with increasing p*

Phase Diagram (Slow Growth), II



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

Note: Phase diagram of "success" / "failure" of HCT coincides with the "possibility" / "impossibility" phase diagram.

Comparison with Bayesian Classifier Threshold (BCT)

Bayesian classifier threshold:

$$t_p = \frac{\beta + r}{2r}\tau_p, \qquad (\tau_p : \text{feature strength in training-}Z\text{-vector})$$

- t_p: 50% selected features are true features
- τ_p: 50% chance for a true feature to be selected
- BCT depends on unknown parameters (β, r)

Surprise: Ideal HCT $\sim \min\{2 \cdot \tau_p,$ Bayes thresh.$\}$

- elevated threshold by the "twice rule"
- HCT is not monotone in feature strength

Comparison with Threshold Choice by Controlling Feature-FDR (FDRT)

For any threshold t,

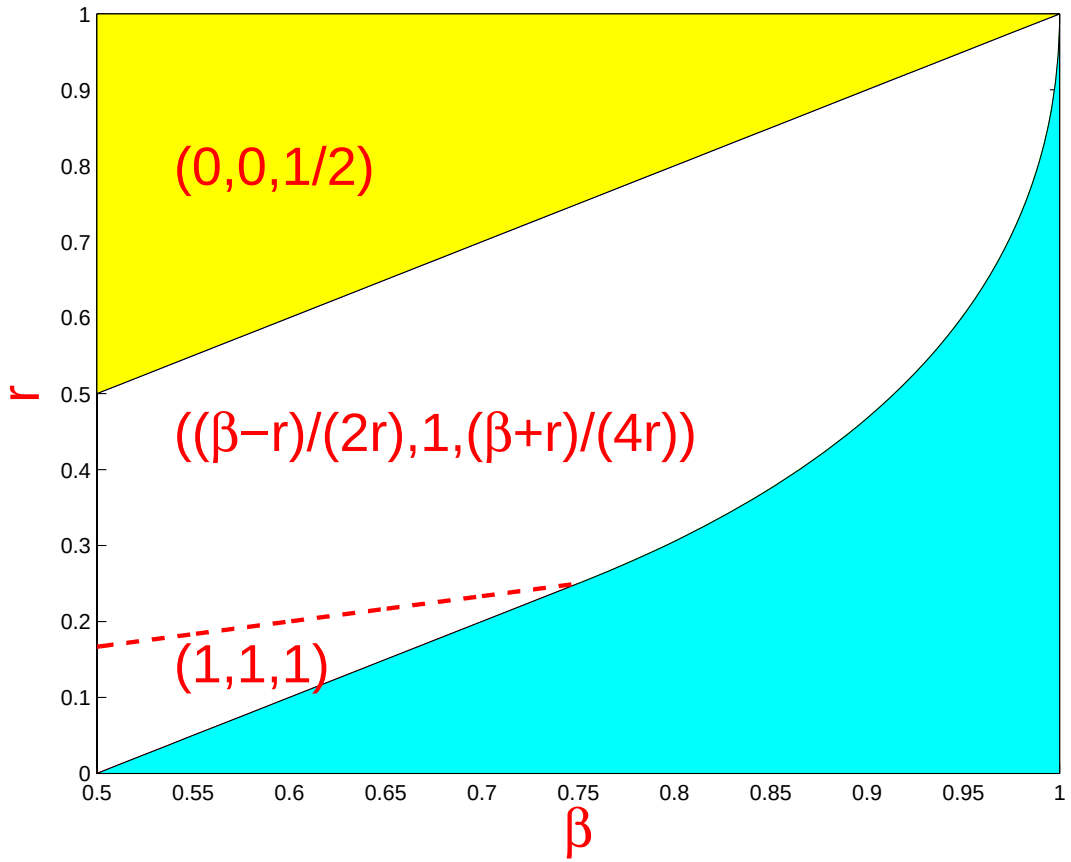
$$\text{Feature-FDR}(t) = \frac{\#\{\text{Falsely Selected Features}\}}{\#\{\text{Total Selected Features}\}}$$

FDRT:

- fix a tolerance parameter $0 < q < 1$ (e.g. $q = .05$)
- FDRT(q): smallest t such that $FDR(t) \leq q$
- similarly, threshold choice by controlling feature-Lfdr, feature-MDR

Challenge: optimal q subtly depends on unknown parameters (β, r)

Phase Diagram of FDR/MDR/Lfdr



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

Three numbers: FDR/MDR/Lfdr

Note "twice rule": $(1 + FDR) = 2 \times Lfdr$

Comparison with Cross-Validation Threshold Choice (CVT)

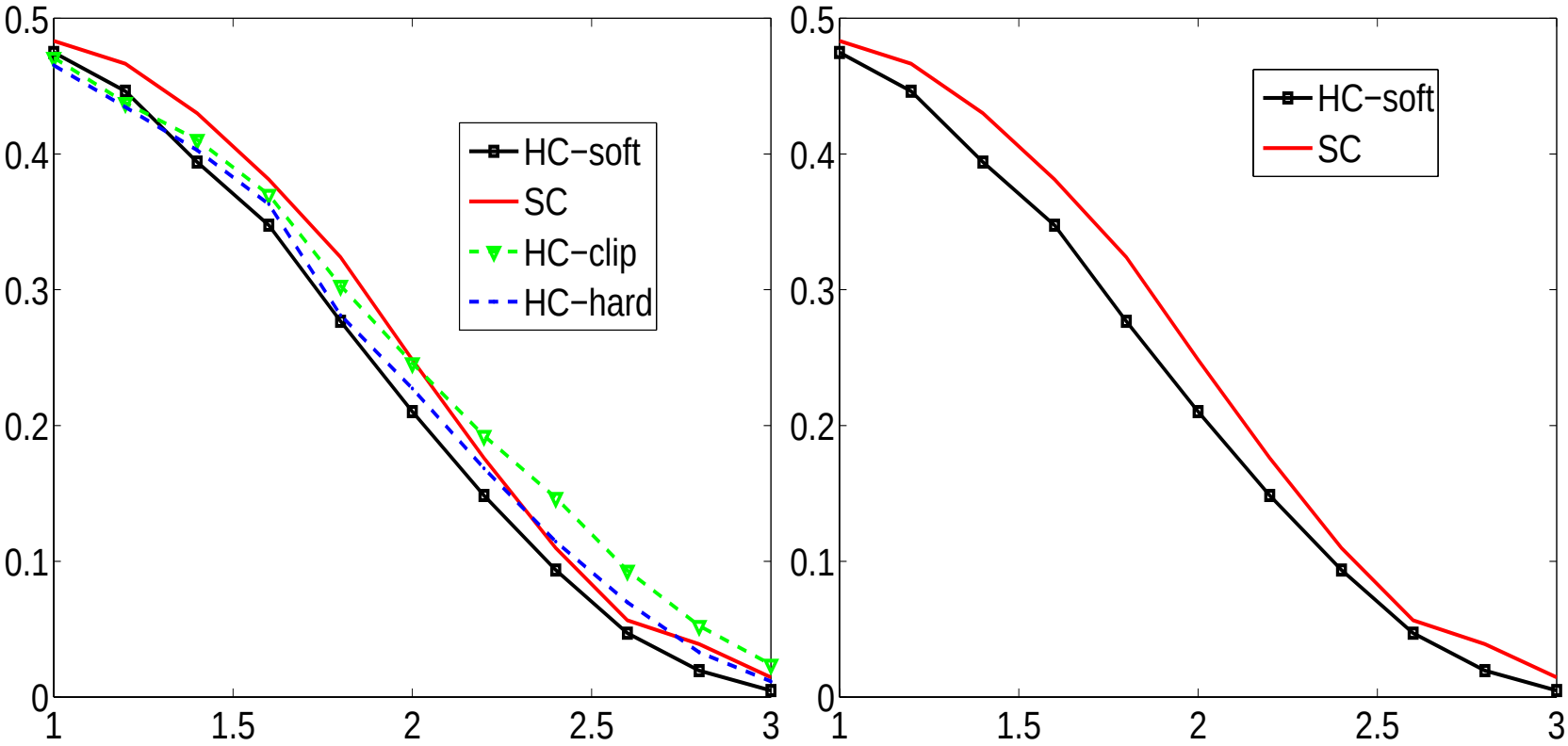
CVT widely used, believed to offer 'optimal threshold' ... but

- Consistency theory does not apply (n small)
- In simulations at representative n,p,
  - **CVT highly variable**
- **CVT computationally much more expensive that HCT**

Ex. Shrunken Centroids (Tibshirani et al. (02))

- Soft Thresholding
- CVT threshold selector

Comparison to Shrunken Centroids (SC)



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

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

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

HCT: Finer Asymptotics

$$\epsilon = p^{-\beta}, \qquad 1/2 < \beta < 1$$

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

$$n = p^{\alpha}$$

In the success region:

- Bias: $\max\{\frac{\tau_p^2}{n},\ p^{-\delta_1(\beta,r,\alpha)}\}$
- Deviation (SD): $p^{-\delta_2(\beta,r,\alpha)}$
- For appropriately small α,

$$\sqrt{MSE} \approx Bias \sim \tau_p^2/n \gg SD$$

Key Technical Result

Fix $(\beta, r) \in (0,1)^2$ with $r > \rho(\beta)$. Let $\epsilon_p = p^{-\beta}$ and $\tau_p = \sqrt{2r\log p}$.
Suppose $n = n_p$ satisfies $\tau_p/\sqrt{n} \to 0$ and $SEP(t_0)/\sqrt{n} \geq 1$.

$$|t_0 - t^*| \leq C \begin{cases} \sqrt{\log p}(\eta_0 + SEP^{-1}(t_0) + \frac{\eta_1(t_0)}{\eta_0}\frac{\tau_p^2}{n}), & \text{Region I,} \\ \frac{1}{\sqrt{\log p}}(\eta_0 + SEP^{-1}(t_0) + \frac{\tau_p^2}{n}), & \text{Region II,} \\ \frac{1}{\sqrt{\log p}}(\eta_1 + \frac{1}{\sqrt{p\eta_0}}(x_0^{-2} + \frac{\eta_0}{\eta_1} + (\tau_p^2/n)^2 + \frac{\tau_p^2}{n}), & \text{Region III.} \end{cases}$$

t_0 : ideal threshold; t^*: optimal threshold

$\eta_0 : \bar{\Phi}(t_0)$; $\eta_1 : \epsilon_p\bar{\Phi}(t_0 - \tau_p)$; $\eta_2 : \epsilon_p\bar{\Phi}(t_0 + \tau_p)$

$x_0 : \sqrt{p} \cdot (\tau_p/\sqrt{n}) \cdot (\eta_1 - \eta_2)/\sqrt{\eta_0 + \eta_1 + \eta_2}$

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: DJ (2008, PNAS): definition, heuristics, practical results
J (2009, PNAS): region of possibility/impossibility
DJ (2009, PTRS-A): phase diagram, first order asymptotics

In preparation: full achievability, extensions, second order asymptotics