New Tools for Sparse Inference: Regime of Moderate Significances

Jiashun Jin

Statistics Department
Purdue University

Collaborators:
David Donoho
Stanford University
Tony Cai and Mark Low
University of Pennsylvania

New Features

- Massive Data
- Sparsity
- Moderately Strong Signals

An Era of Massive Data

- Proteomic: head-neck cancer
  - 221 patients
  - $33,000$ data points for each
- Sports: baseball player dataset
  - 80 baseball players
  - 10^6 data points for each
- Image Analysis

Sparsity

- A natural phenomenon found in many application areas
- Only **a small fraction** of the data contains non-null effect or **signals**, others are null effect or noise
- Many seemingly intractable statistical problems can be successfully attacked simply by **exploiting sparsity**

Problems in Sparse Inferences

- Whether or not: whether there is any signal or not
- How many
- Where

Two Types of Signal

1. Very strong signal:
   - stand out for themselves
   - relatively easier to tell "where", e.g. thresholding
2. Moderately strong signal:
   - larger than typical noise, but not larger than all of them
   - not strong enough to stand out for themselves
   - can't tell "where"

Example: hidden sparse "spikes" in white noise

- noise: n *iid* samples from $N(0,1)$
- extreme noise: $\approx \sqrt{2\log n}$
- moderately significant signal: e.g. $\approx \sqrt{\log n}$.

Regime of Moderate Significances

- Found in many applications
- Statistically a field only in its infancy
- Poses new challenges in data analysis:
  - tiny in fraction: averaging won't work
  - moderately strong: thresholding won't work (extreme values are noise!)
- Needs new formulations and new tools

Example I: Covert Communication

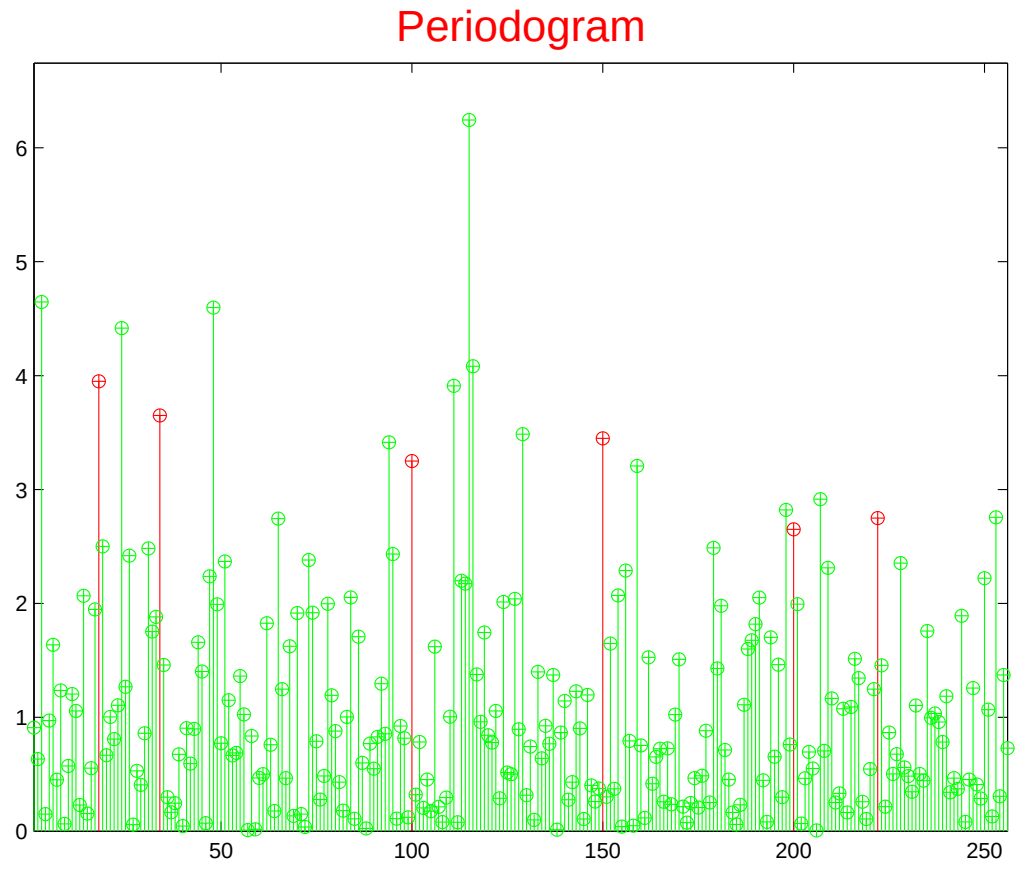


Figure 1: Covert Communication, 6 **red** lines illustrate the spectrum used to send **signals**, 250 green lines illustrate exponential noise.

Example II: Kuiper Belt Object

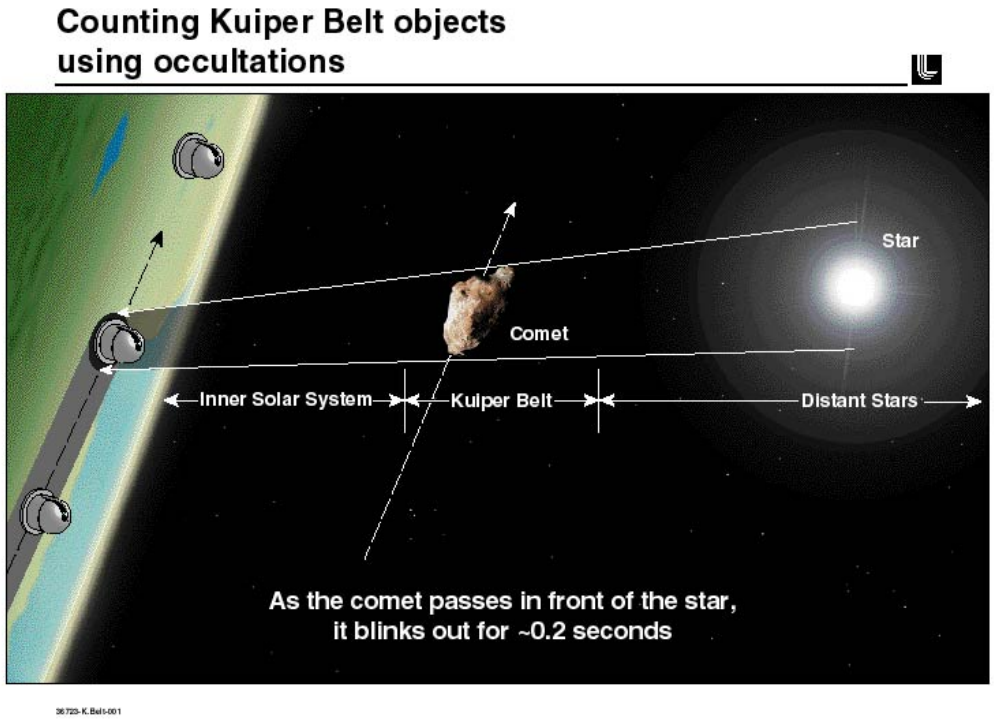


Figure 2: Taiwanese-American Occultation Survey. $10^{10} - 10^{12}$ tests per year, only tens or hundreds contain Kuiper Belt Object (KBO).

Abstraction

- Subtle effect **sparsely scattered** across the observations
- Individual effects not strong enough to stand out, i.e. **moderately significant**
- Particularly interested in the proportion:

$$\epsilon_n = \frac{\#\{\text{data point containing a signal}\}}{n}$$

Two inter-connected Problems:

- *whether or not*: testing $\epsilon_n = 0$ vs. $\epsilon_n > 0$
- *how many*: estimate ϵ_n
- signal not strong enough to tell "where"

Agenda

1. Testing $\epsilon_n = 0$ vs. $\epsilon_n > 0$
   - Detecting Sparse Gaussian mixtures
   - Optimal adaptivity of Higher Criticism (HC^*)
2. Estimating ϵ_n
   - MR lower bound (Meinshausen and Rice)
   - CJL lower bound (Cai, Jin, and Low)

Part I. Testing $\epsilon_n = 0$ vs. $\epsilon_n > 0$

Ingster (1997,1999), Jin(2003, 2004)

Hypothesis Testing:

$$H_0: \ X_i \overset{i.i.d}{\sim} N(0,1), \qquad 1 \le i \le n,$$

$$H_1^{(n)}: \ X_i \overset{i.i.d}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_n, 1), \qquad 1 \le i \le n.$$

Goal: delineate for what (ϵ_n, μ_n) H_0 and $H_1^{(n)}$ **separable asymptotically**

The (β, r)-plane

Calibrate with:

$$\epsilon_n = n^{-\beta}, \qquad \underline{0.5 < \beta < 1}$$

$$\mu_n = \sqrt{2r \log n}, \qquad \underline{0 < r < 1}$$

1. Detection problem with (β, r) in this range is **new** and **challenging**
2. **Subtlety** of the problem:
   - $\epsilon_n \ll \frac{1}{\sqrt{n}}$: **very small** fraction non-zeros means

     $$e.g. \qquad \sqrt{n}\bar{X}_n \text{ would fail}$$

   - $\mu_n < \sqrt{2\log n}$: signals of **only moderate** significance

     $\sqrt{2\log n} \approx$ largest X_i from "true null" component hypotheses

Detection Boundary

Theorem 1. (*Ingster 1999, Jin 2004*). If $\epsilon_n = n^{-\beta}$, $\mu_n = \sqrt{2r\log n}$, $\frac{1}{2} < \beta < 1$, and $0 < r < 1$, then:

$$\text{If } r > \rho(\beta), \quad H_0 \text{ and } H_1^{(n)} \text{ separate asymptotically,}$$

$$\text{If } r < \rho(\beta), \quad H_0 \text{ and } H_1^{(n)} \text{ merge asymptotically.}$$

We call $r = \rho(\beta)$ the "detection boundary":

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

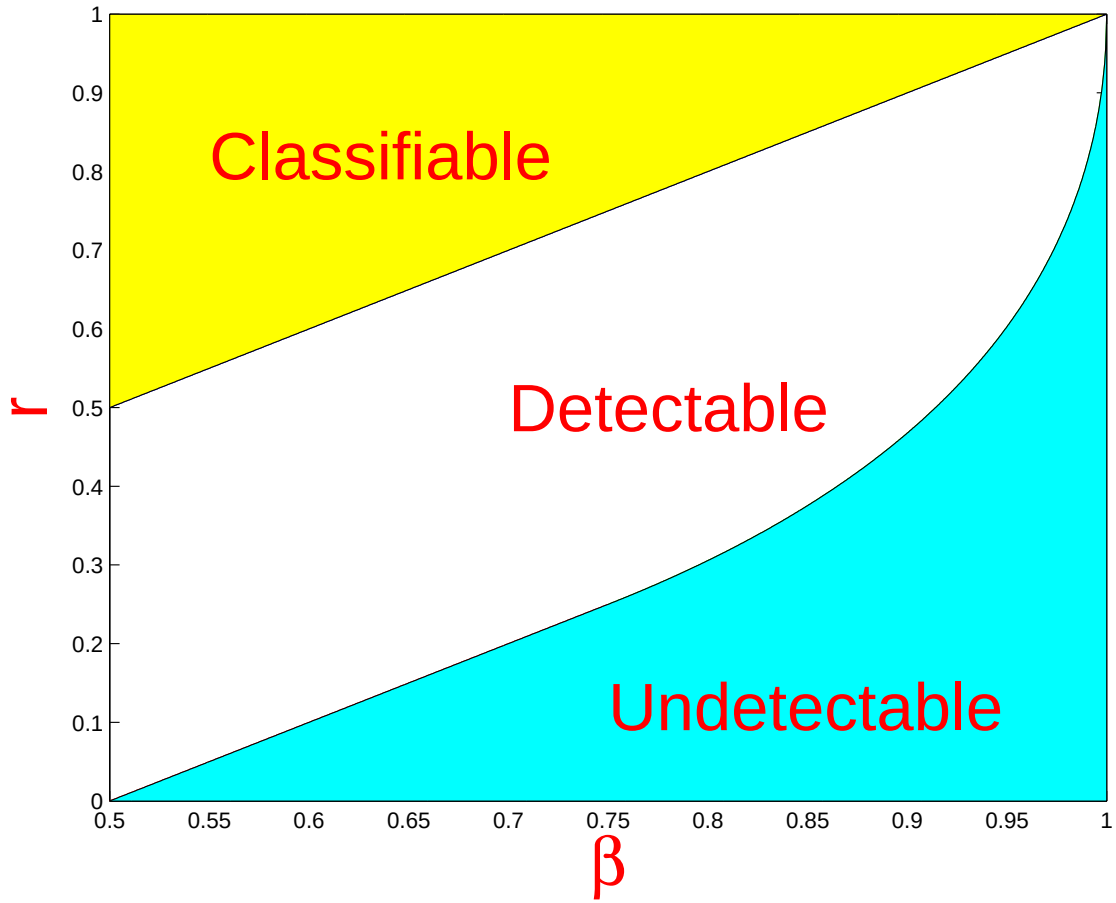


Figure 3: Regions in (β, r) plane. "Detection boundary" separates green and white regions. In yellow region, possible to isolate signals from noise (FDR thresholding).

Ideas for the Proof

Recall

$$\epsilon_n = n^{-\beta}, \qquad \underline{0.5 < \beta < 1}$$

$$\mu_n = \sqrt{2r\log n}, \qquad \underline{0 < r < 1}$$

- For fixed (β, r), Neyman-Pearson test is optimal
- Key: when (β, r) falls **exactly on** the detection boudary

$$LR_n^* \longrightarrow_{weakly} \begin{cases} \nu_0, & \text{under null} \\ \nu_1, & \text{under alternative} \end{cases}$$

$r = \rho(\beta)$	$\frac{1}{2} < \beta < \frac{3}{4}$	$\beta = \frac{3}{4}$	$\frac{3}{4} < \beta < 1$
ν_0	$N(-\frac{1}{2}, 1)$	$N(-\frac{1}{4}, \frac{1}{2})$	an I.D. Law
ν_1	$N(\frac{1}{2}, 1)$	$N(\frac{1}{4}, \frac{1}{2})$	an I.D. Law

Where's the Evidence Against H_0?

Recall

$$\mu_n = \sqrt{2r \log n}, \qquad 0 < r < 1$$

- The likelihood ratio peaks near

$$x_n = \min\{2\mu_n, \sqrt{2 \log n}\}$$

- Interestingly, **not** near μ_n
- Position of peak depends on (β, r)

Limitation of likelihood ratio:

- Attached to Gaussian mixture model
- Particularly, not adaptive to unknown (β, r)

Higher Criticism Statistic: HC^*

Denote individual p-value: $p_i = P\{N(0,1) \geq X_i\}$

1. Sort p-values: $p_{(1)} < p_{(2)} \ldots < p_{(n)}$
2. Calculate i^{th} z-score:

$$HC_{n,i} = \sqrt{n}\left[\frac{\frac{i}{n} - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}\right]$$

3. Take maximum:

$$HC_n^* = \max_{\{1 \leq i \leq n\}} HC_{n,i}$$

4. Look for unusually large amount of "moderate significances"

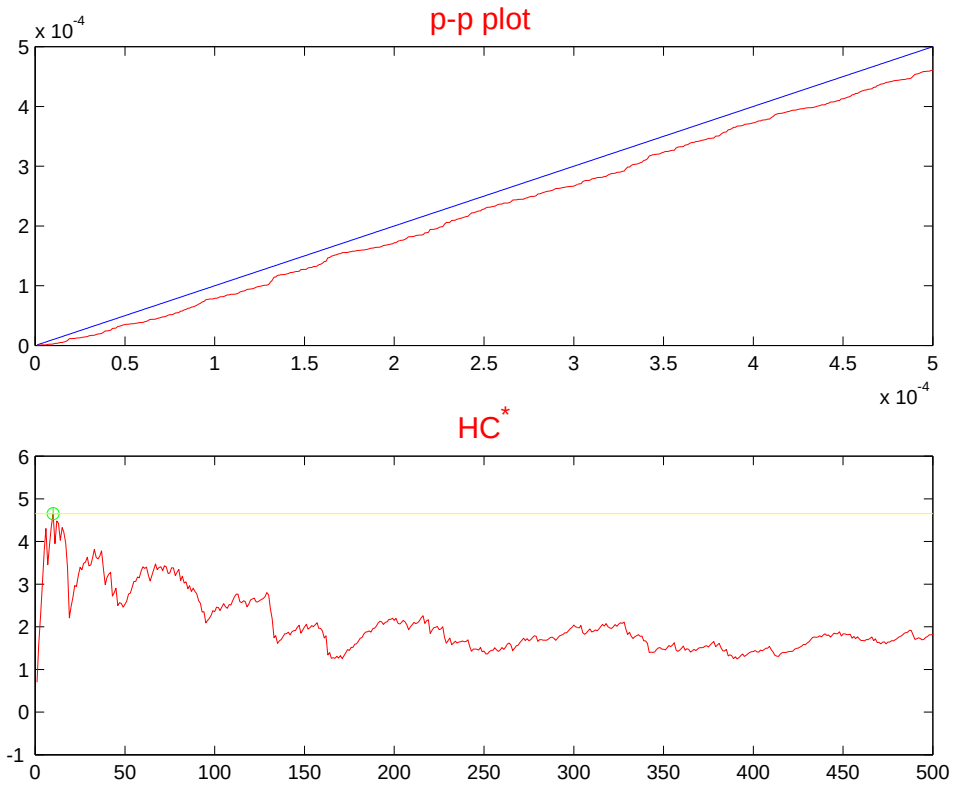


Figure 4: p-p plot (**top**) and z-scores versus p-values (**bottom**). $HC^* \approx 4.7$ is the maximum of z-scores, which is associated with the 11^{th} smallest p-value (**green dot**).

How to use Higher Criticism

To use Higher Criticism for *testing*, we need to determine a critical value $h(n, \alpha)$, so

$$\text{Reject } H_0 \text{ if and only if} \quad HC_n^* \geq h(n, \alpha)$$

where

$$P_{H_0}\{HC_n^* \geq h(n, \alpha)\} = \alpha$$

Theorem 2. Under H_0,

$$\frac{HC_n^*}{\sqrt{2 \log\log n}} \quad \longrightarrow_p \quad 1, \qquad n \to \infty$$

Proof: use Erodös-Kac Theorem

Formal Specification of Critical Value

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

$$P_{H_0}\{HC_n^* \geq h(n,\alpha)\} = \alpha$$

Implication of Theorem 2: for moderate α, we expect

$$h(n,\alpha) \sim \sqrt{2\log\log n}$$

We say $\alpha_n \to 0$ *slowly enough* if

$$\frac{h(n,\alpha_n)}{\sqrt{2\log\log n}} \quad \to \quad 1, \qquad n \to \infty$$

Optimal Adaptivity of Higher Criticism

Donoho and Jin (2004). Ann. Statist.

Theorem 3. Consider the Higher Criticism test that rejects H_0 when

$$HC_n^* \geq h(n, \alpha_n)$$

where the level $\alpha_n \to 0$ *slowly enough*. For every alternative $H_1^{(n)}(r, \beta)$ where r exceeds the detection boundary $\rho(\beta)$ — so that the likelihood ratio test (LRT) would have full power — Higher Criticism test also has full power:

$$P_{H_1^{(n)}}\{\text{Reject } \ H_0\} \ \ \to \ \ 1.$$

Intuition: HC_n^* is able to adaptively "find the evidence" near $x_n = \min\{2\mu_n, \sqrt{2\log n}\}$

Comparison to Other Statistics

- Kendall & Kendall's pontogram
- Berk-Jones (1978)

Part II. Estimating ϵ_n

Sparse Gaussian mixture model:

$$(X_i|\mu_i) \sim (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_i, 1), \qquad \mu_i \overset{iid}{\sim} H_n, \qquad 1 \le i \le n$$

- Bayesian version of "needles in haystack"
- H: distribution of signals
  - Include previous two-point mixture as a special case
  - $H = H_n$: triangle array
  - For simplicity: $P\{H > 0\} = 1$

Possible and Impossible

Without further constraint on $H = H_n$, the possibility of $\epsilon_n = 1$ can never be ruled out:

- Non-trivial confidence upper bound is impossible
- Consistent estimate is impossible
- However, non-trivial confidence lower bound is possible

Goal

1. For any given level $0 < \alpha < 1$, construct a lower bound $\hat{\epsilon}_n$ which holds uniformly for all such Gaussian mixtures:

$$P\{\hat{\epsilon}_n \leq \epsilon_n\} \qquad \geq \qquad (1-\alpha)$$

2. View the lower bound as an estimator of ϵ_n and study the consistency for more specified models

Meinshausen and Rice's Lower Bound

A family of lower bounds indexed by $0 < \alpha < 1$:

$$\hat{\epsilon}_n^{MR} = \sup_t \left\{ \frac{\Phi(t) - F_n(t) - (h(n,\alpha)/\sqrt{n}) \cdot \sqrt{\Phi(t)(1-\Phi(t))}}{\Phi(t)} \right\}$$

Notations:

- $\Phi(t)$: CDF of $N(0,1)$
- $F_n(t)$: empirical CDF
- $h(n,\alpha)$: critical value of Higher Criticism

$$P_{H_0}\{HC_n^* \geq h(n,\alpha)\} = \alpha$$

Uniform Lower Bound

Meinshausen and Rice's construction:

- uniformly honest:

$$P\{\hat{\epsilon}_n^{MR} \leq \epsilon_n\} \quad \geq \quad (1-\alpha)$$

- valid even when the signals are not from Gaussian mixtures

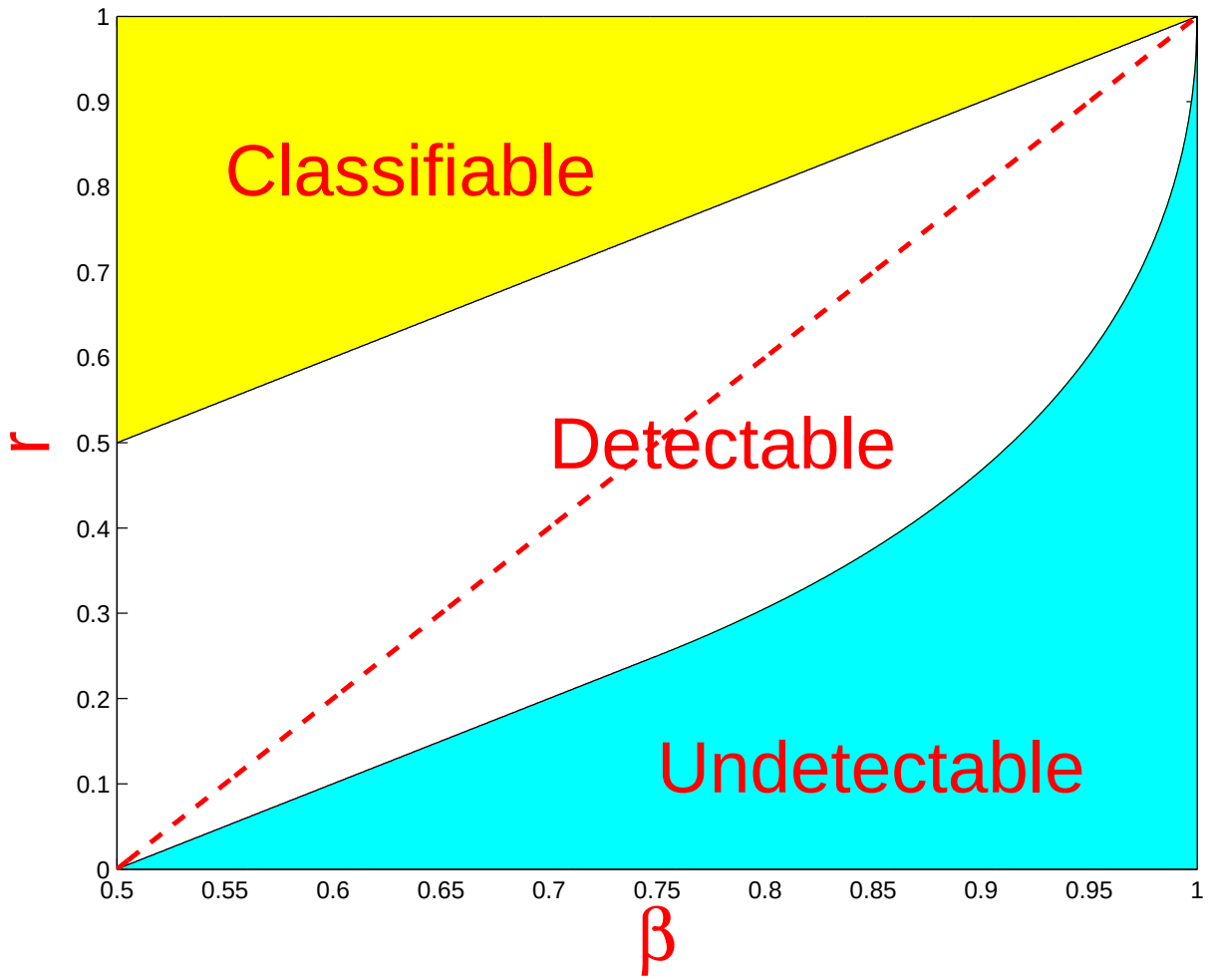
Consistency for Two-point Mixtures

Theorem 4. (*Meinshausen and Rice 2004, to appear in Ann. Statist.*). If $\epsilon_n = n^{-\beta}$, $\mu_n = \sqrt{2r\log n}$, $\frac{1}{2} < \beta < 1$, $0 < r < 1$, and $\alpha_n \to 0$ slowly enough then:

$$\lim_{n\to\infty}\left[\frac{\hat{\epsilon}_n^{MR}}{\epsilon_n}\right] = \begin{cases} 1, & r > 2\beta - 1, \\ 0, & r < 2\beta - 1. \end{cases}$$

Recall that α_n tends to 0 slowly enough means:

$$\frac{h(n,\alpha_n)}{\sqrt{2\log\log n}} \quad \to \quad 1, \quad n \to \infty$$


Classifiable
Detectable
Undetectable
r
β
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

Unanswered Questions

The lower bound:

- only consistent in a region **smaller** than the detectable region
- they conjecture: "...it is clear that it is somewhat easier to test for the global null hypothesis than to estimate the proportion"

Unanswered Question: the precise region over which consistent estimation of ϵ_n is possible?

Answer: surprisingly, it coincides with the detectable region!

CJL Construction of Lower Bound

Cai, Jin, Low 2005, manuscript

1. Construct a confidence envelop for the underlying cdf $F(t)$:

$$F^{\pm}(t) = \frac{2F_n(t) \pm (h/\sqrt{n}) \cdot \sqrt{h^2/n + (4F_n(t) - 4F_n^2(t))} + h^2/n}{2(1 + h^2/n)},$$

where $h = h(n, \alpha)$ is the critical value of Higher Criticism:

$$P_{H_0}\{HC_n^* \geq h(n, \alpha)\} = \alpha$$

2. Lay out an equal-spaced grid:

$$t_j = \frac{(j-1)}{\sqrt{2\log n}}, \qquad 1 \leq j \leq (2\log n + 1)$$

3. For adjacent pair of grid points, using $F^{\pm}$ to construct lower bounds $\hat{\epsilon}_n^{(j)}$
   - Solving a bridging quantity:
   $$D(\hat{\mu}_j) = \frac{\Phi(t_j) - F^+(t_j)}{\Phi(t_{j+1}) - F^-(t_{j+1})}, \quad D(\mu) \equiv \frac{\Phi(t_j) - \Phi(t_j - \mu)}{\Phi(t_{j+1}) - \Phi(t_{j+1} - \mu)}$$
   - Plug in:
   $$\hat{\epsilon}_n^{(j)} = \begin{cases} \frac{\Phi(t_j) - F^+(t_j)}{\Phi(t_j) - \Phi(t_j - \hat{\mu}_j)}, & \text{above solution exists} \\ 0, & \text{otherwise} \end{cases}$$
   - Each $\hat{\epsilon}_n^{(j)}$ underestimates $\hat{\epsilon}_n^{(j)}$, most *seriously*, some *slightly*
4. Take supremum:
   $$\hat{\epsilon}_n^{CJL} = \sup_j \hat{\epsilon}_n^{(j)}$$

Uniform Lower Bound

Let $h(n,\alpha)$ be the critical value of Higher Criticism:

$$P_{H_0}\{HC_n^* \geq h(n,\alpha)\} = \alpha,$$

then

$$P\{\epsilon_n^{CJL} \leq \epsilon_n\} \quad \geq \quad (1-\alpha),$$

uniformly for all Gaussian mixtures with cdf:

$$F(t) = (1-\epsilon_n)\Phi(t) + \epsilon_n \int \Phi(t-\mu)dH_n, \qquad P\{H_n > 0\} = 1$$

Optimally Consistent for Two-point Mixtures

Theorem 5. (*Cai, Jin, Low 2005, manuscript*). If $\epsilon_n = n^{-\beta}$, $\mu_n = \sqrt{2r\log n}$, $\frac{1}{2} < \beta < 1$, $0 < r < 1$. Suppose $\alpha_n \to 0$ slowly enough, then for every (r, β) where r exceeds the detection boundary, the CJL lower bound with $h(n, \alpha_n)$ estimates ϵ_n consistently.

Recall that α_n tends to 0 slowly enough means:

$$\frac{h(n,\alpha_n)}{\sqrt{2\log\log n}} \quad \to \quad 1, \qquad n \to \infty$$

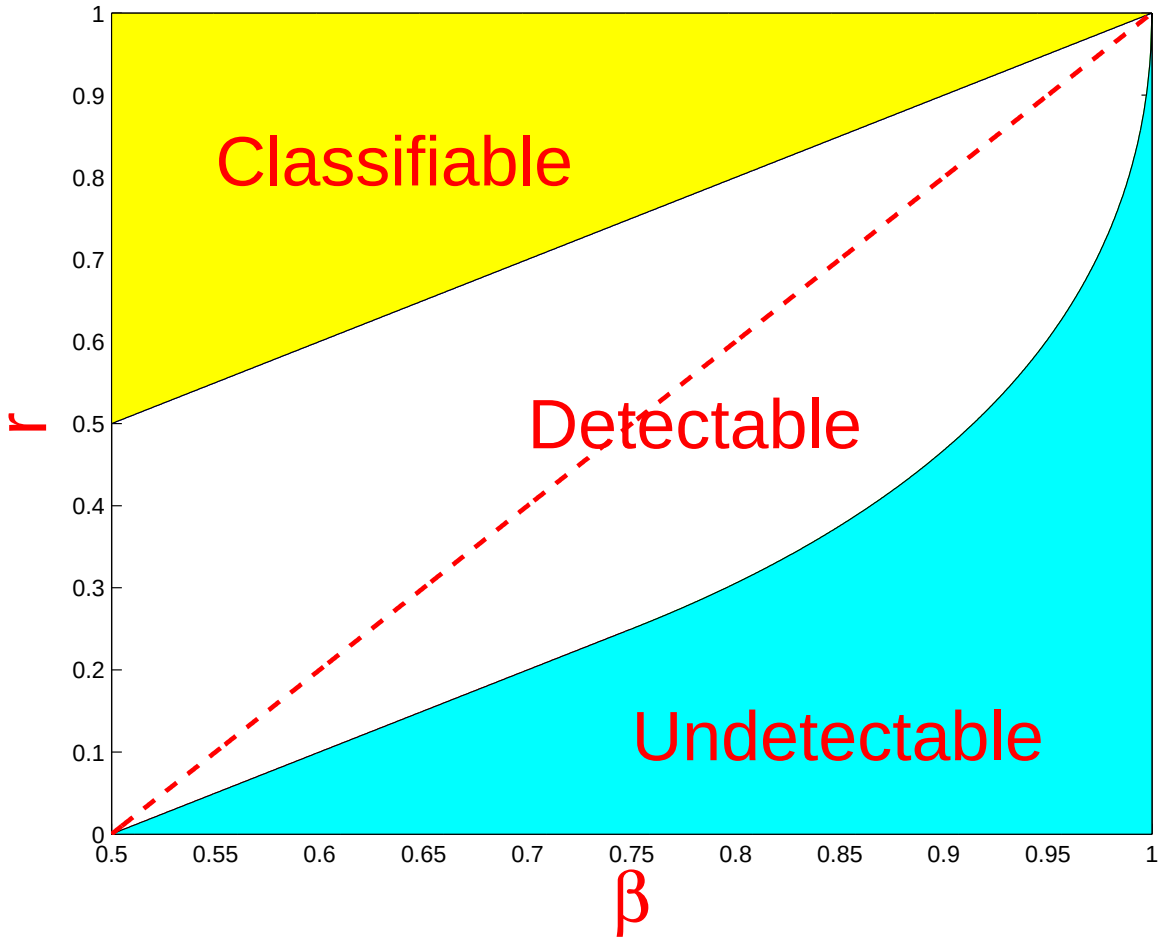


Figure 5: The white region is a *brain teaser*: even when (r, β) are completely known, it is impossible to tell "where"

Obtains Optimal Rate in Mean Square Error

Theorem 6. (*Cai, Jin, Low 2005, manuscript*). If $\epsilon_n = n^{-\beta}$, $\mu_n = \sqrt{2r\log n}$, $\frac{1}{2} < \beta < 1$, $0 < r < 1$. Suppose we take $h(n,\alpha) = 4\sqrt{2\pi}\log^{3/2}(n)$ in the construction of the lower bound, then for every (r,β) where r exceeds the detection boundary:

$$E\left[\frac{\hat{\epsilon}_n^{CJL}}{\epsilon_n} - 1\right]^2 \qquad \leq \qquad L_n \cdot \inf_{\hat{\epsilon}_n}\left\{E\left[\frac{\hat{\epsilon}_n}{\epsilon_n} - 1\right]^2\right\}$$

where L_n is a generic multi-$\log(n)$ terms.

Comparisons of Three Procedures

	Higher Criticism	MR Lower Bound	CJL Lower Bound
Test	✓		
Estimate		✓	✓
Valid Beyond Gaussian	✓	✓	
Optimally Adaptive	✓		✓
Optimal in MSE			✓
Not Heavy-tailed			✓

Take Home Messages

1. Introduced and compared three new procedures
2. Found precise demarcations for:
   - when it is possible to tell $\epsilon_n > 0$
   - when it is possible to consistently estimate ϵ_n
   - when it is possible to tell "where"
3. Proved optimality of Higher Criticism and CJL lower bound
4. Surprisingly, whenever it is able to tell $\epsilon_n > 0$, it is also possible to consistently estimate it

See my website:

www.stat.purdue.edu/~ jinj

for applications in Cosmology and Astronomy

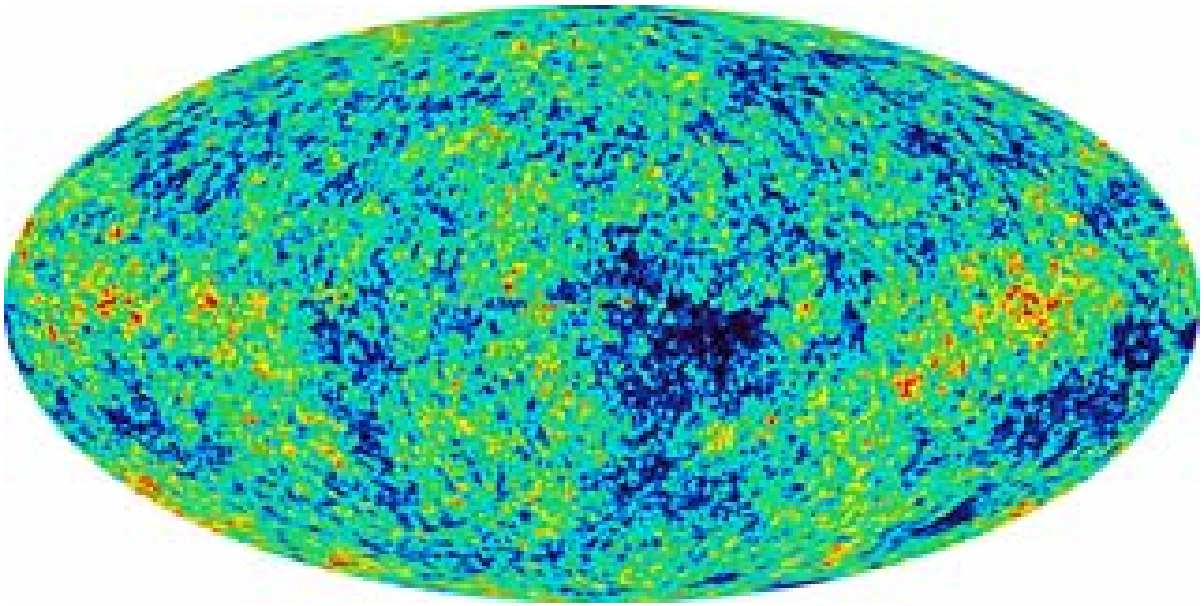


Figure 6: Wilkson Microwave Anisotropy Project (WMAP).

Acknowledgements

David Donoho	Collaboration Discussion Encouragements	Stanford University
Betsy Becker	Discussion	Michigan State U.
Yoav Benjamini	References	Tel Aviv University
Henry Braun		Princeton University
Larry Brown		U. Pennsylvania
Iain Johnstone		Stanford University
Eric Kolaczyk		Boston University
Art Owen		Stanford University
John Rice		UC Berkeley
David Siegmund		Stanford University
Howard Wainer		ETS

Simulations

- $3,500$ independent cycles of simulations
- in each cycle:
  - draw n samples from $(1-\epsilon)N(0,1) + \epsilon N(\mu, 1)$
  - pick $\alpha = 0.5\%$, 5%, and 25%
  - for each α, calculate ratios $(\hat{\epsilon}_n^{MR}/\epsilon_n)$ and $(\hat{\epsilon}_n^{*}/\epsilon_n)$
- where
  - $n = 10^7$
  - $\epsilon_n = 10^{-4}$
  - $\mu \approx 4$

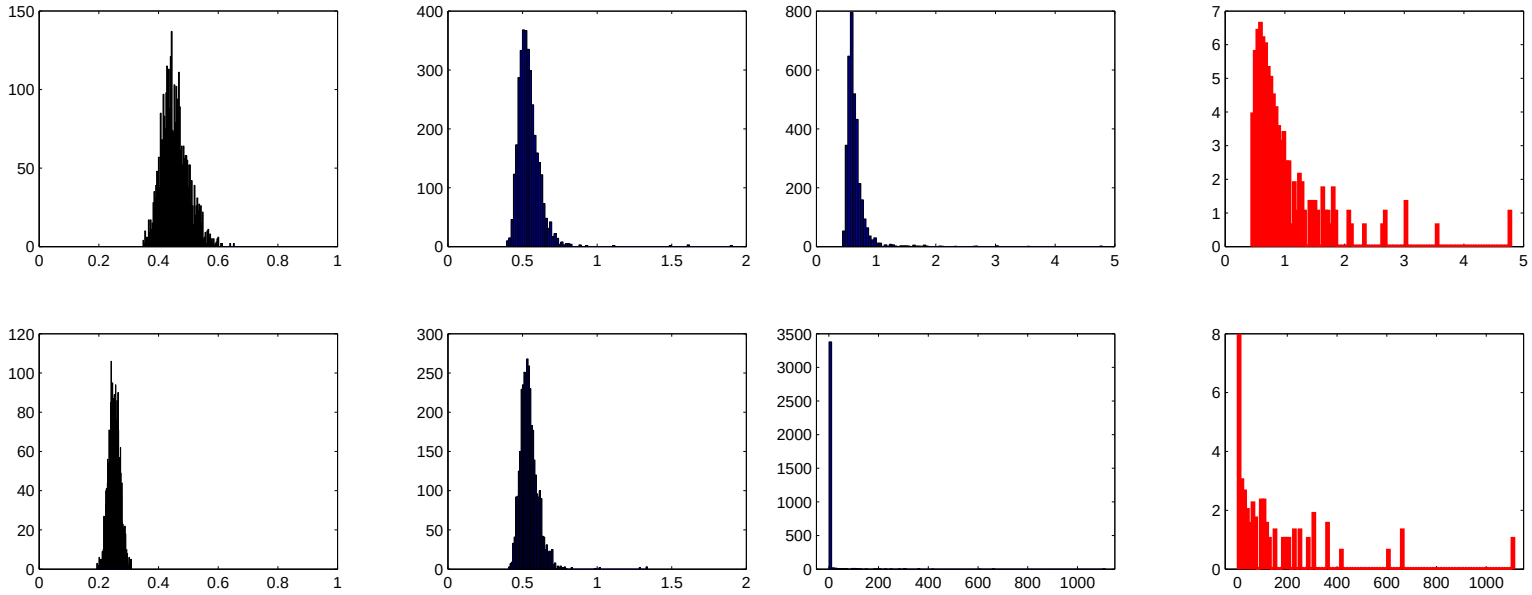


Figure 7: Blue: histograms of ratios correspond to $\alpha = 0.5\%$, 5%, and 25%. Column 4 is log-histogram corresponds to Column 3. Top: our lower bound. Bottom: Meinshausen and Rice's lower bound.