Proportion of Non-zero Normal Means: Oracle Estimators and Applications to CGH Array

Jiashun Jin

Statistics Department
Purdue University

Collaborators

Alphabetically:

Tony Cai	University of Pennsylvania
Jie Peng	University of California at Davis
Pei Wang	Fred Hutchinson Cancer Research Center

Massive Data

- Massive investment in data collection/processing, many areas of science and business
- Massive datasets routinely generated:
  - Genomics and proteomics
  - Cosmology and astronomy
  - Financial tick-by-tick data
  - fMRI

High-dimensional Data Analysis

- Traditional statistical data
  - Sample: human being
  - Dimension: blood pressure, weight, height ....
  - Ex. 20 samples, 3 dimensions
- Modern statistical data
  - Sample: human being
  - Dimension: vectors, curves, spectra, images ...
  - Ex. 100 samples, $10,000$ dimensions

Large-scale multiple hypothesis testing

1. Many *null* hypotheses:

$$H_1, H_2, \ldots, H_n$$

2. Many test statistics (summary statistics, regression coefficients, transform coefficients):

$$X_1, X_2, \ldots, X_n$$

Terminology:

- If H_j is true, call X_j a null effect (noise, haystack)
- Otherwise, call X_j a non-null effect (signal, needle)

Two Types of Signal

1. Very strong signal:
   - stand out for themselves
   - relatively easier to tell "where", e.g. thresholding
   - relatively few in numbers
2. Moderately strong signal:
   - not strong enough to stand out
   - can't be isolated or detected **individually**
   - dominating in numbers

For Today

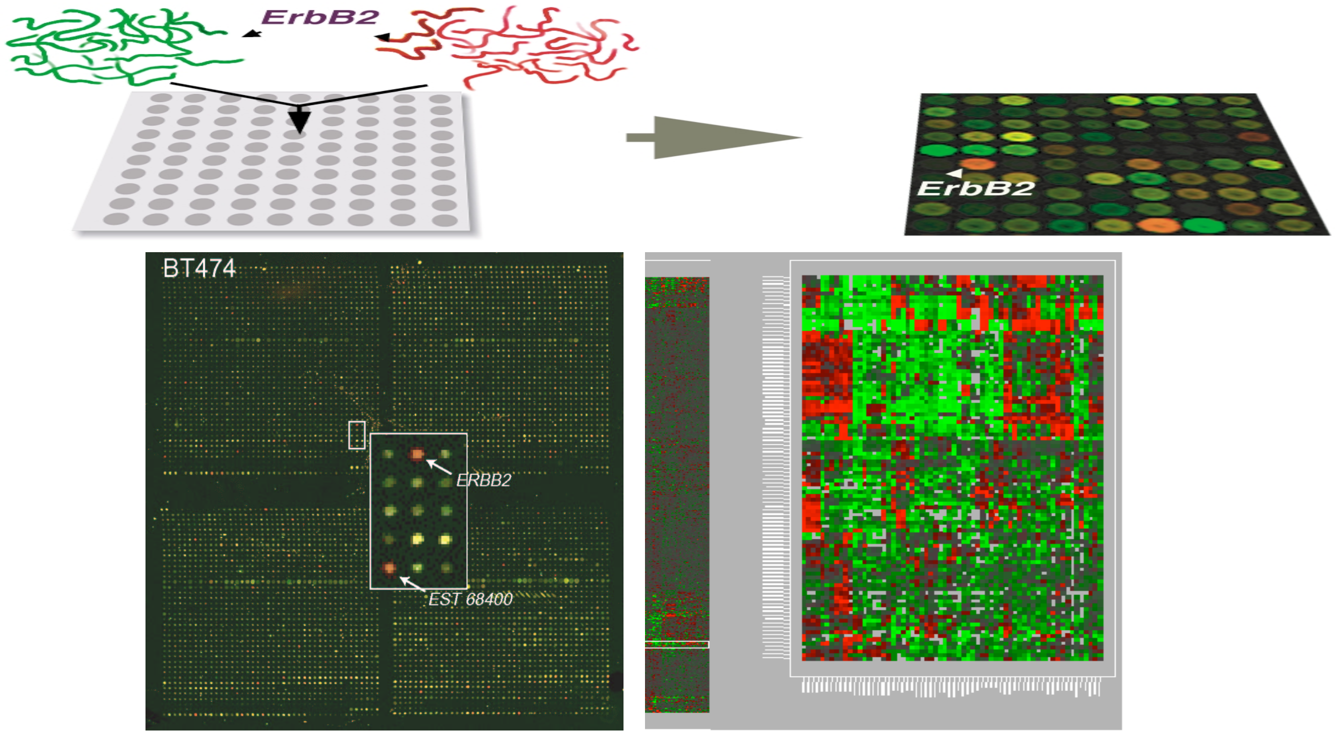
Estimating the **proportion of signals**:

$$\epsilon_n = \frac{\#\{j : H_j \text{ is untrue}\}}{n}$$

Focusing on **faint/moderately strong** signals:

- Signals not strong enough to be isolated *individually*
- Still possible to estimate the proportion

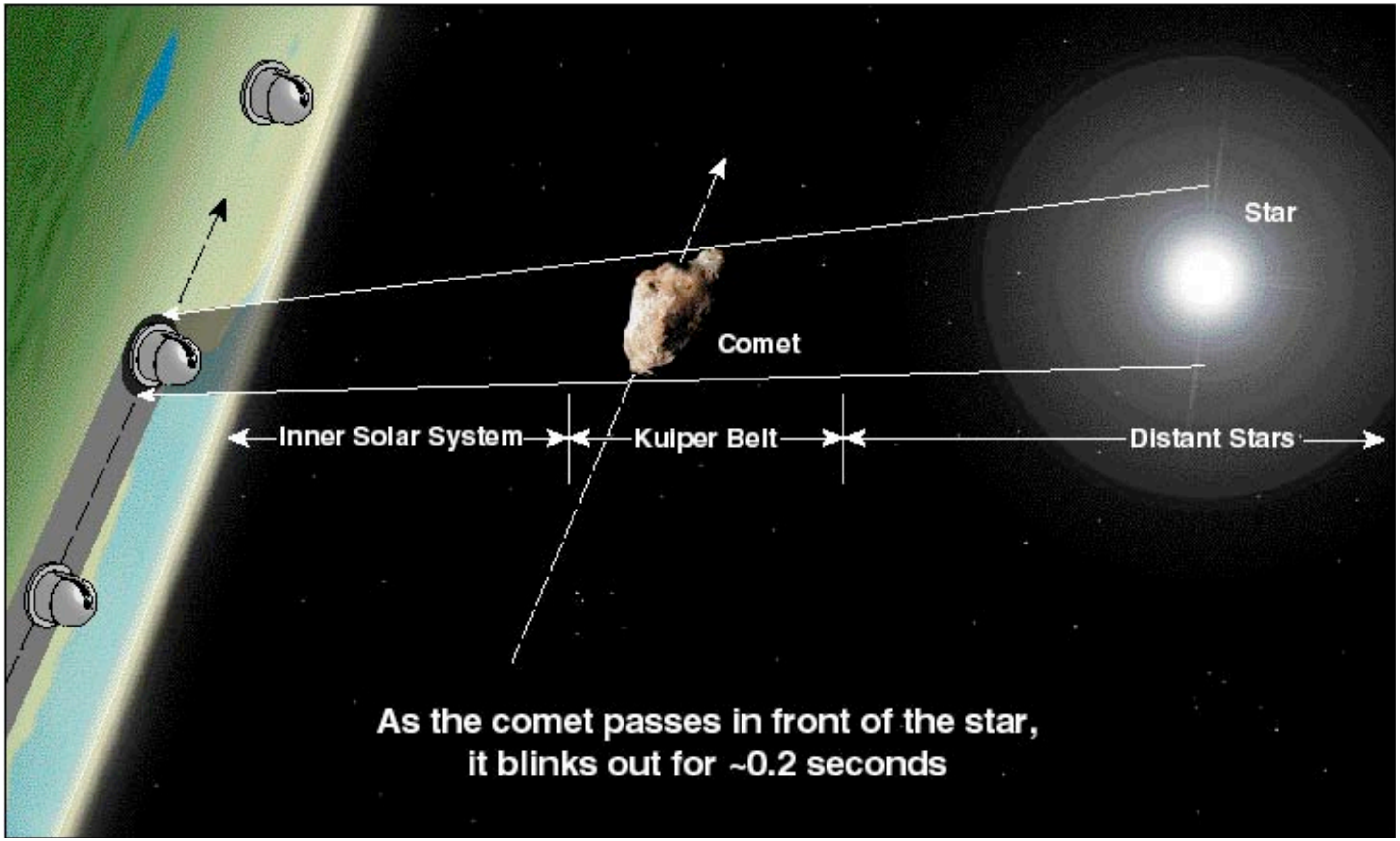
Example I: Lung Cancer CGH Array



Paired CGH profile (left) and mRNA profile (right)

CGH: Comparative Genomic Hybridization.

Example II: Kuiper Belt Object (KBO)



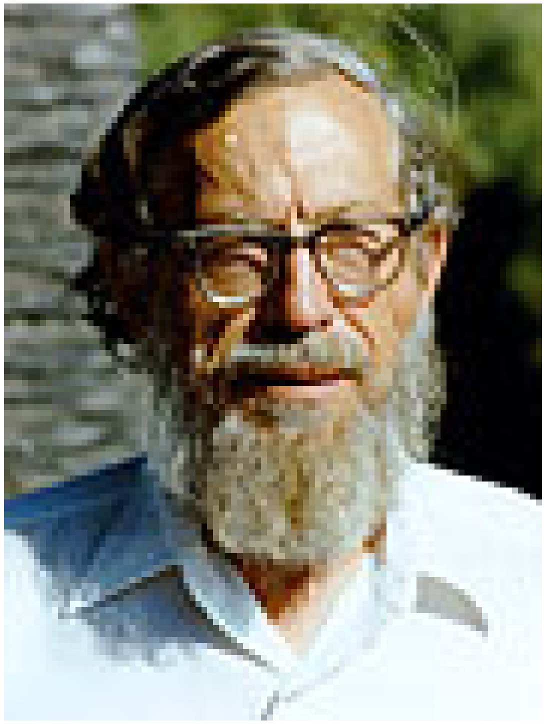
Taiwanese-American Occultation Survey (TAOS)

$10^{10} - 10^{12}$ tests, only tens or hundreds contain KBO

Agenda

1. Proportion of nonzero normal means:
   - Universal oracle equivalence
   - Uniformly consistent estimators
   - Extensions to heteroscedastic models
2. Comparison with other approaches
3. Applications to CGH array

Stein's n-normal Means Setting



Charles Stein

- n data points, n parameters:

$$X_j = \mu_j + \epsilon_j, \quad \epsilon_j \overset{iid}{\sim} N(0,1)$$

- A snapshot of an n-vector
- Caught a lot of enthusiasm
  - captures the essence of "high dimension" data
  - handle many applications
  - tractable

Estimating n-normal Means

Goal: Estimating μ_j's *simultaneously*

$$X_j \sim \mu_j + \epsilon_j, \qquad \epsilon_j \overset{iid}{\sim} N(0,1), \qquad j = 1, \dots, n$$

- MLE: $\hat{\mu}_j = X_j$
- Stein's shrinkage
- Wavelets and non-parametric estimation:
  - $Y(t) = f(t) + W(t), \;\; 0 < t < 1$
  - X_j: WC of $Y(t)$. WC: wavelet coefficients
  - μ_j: WC of $f(t)$
  - ϵ_j: WC of $W(t)$

Testing n-normal Means

- n test statistics

$$X_j \sim \mu_j + \epsilon_j, \qquad \epsilon_j \overset{iid}{\sim} N(0,1), \qquad j = 1, \ldots, n$$

- n hypotheses

$$\begin{aligned} \mu_j &= 0, \qquad \text{if } H_j \text{ is true} \\ \mu_j &\neq 0, \qquad \text{if } H_j \text{ is untrue} \end{aligned}$$

- An insurgence of research interest
  - Driven by development in multiple testing and microarray
  - Bridge for understanding more complicated models

Proportion of nonzero Normal Means

Jin (2006), under review

$$X_j \sim \mu_j + \epsilon_j, \qquad \epsilon_j \overset{iid}{\sim} N(0,1), \qquad j = 1, \ldots, n$$

Goal:

- Estimating the proportion of nonzero normal means

$$\epsilon_n(\mu) = \frac{\#\{j : \mu_j \neq 0\}}{n}$$

- Focusing on faint signals (i.e. μ_j are small)

Where is the Information?

1. Tukey's wisdom: which part of the data contains the information?

2. Where is the *information of the proportion?*

   - Surprisingly, not in the spatial domain (densities, cdfs, moments, data tails, etc.)

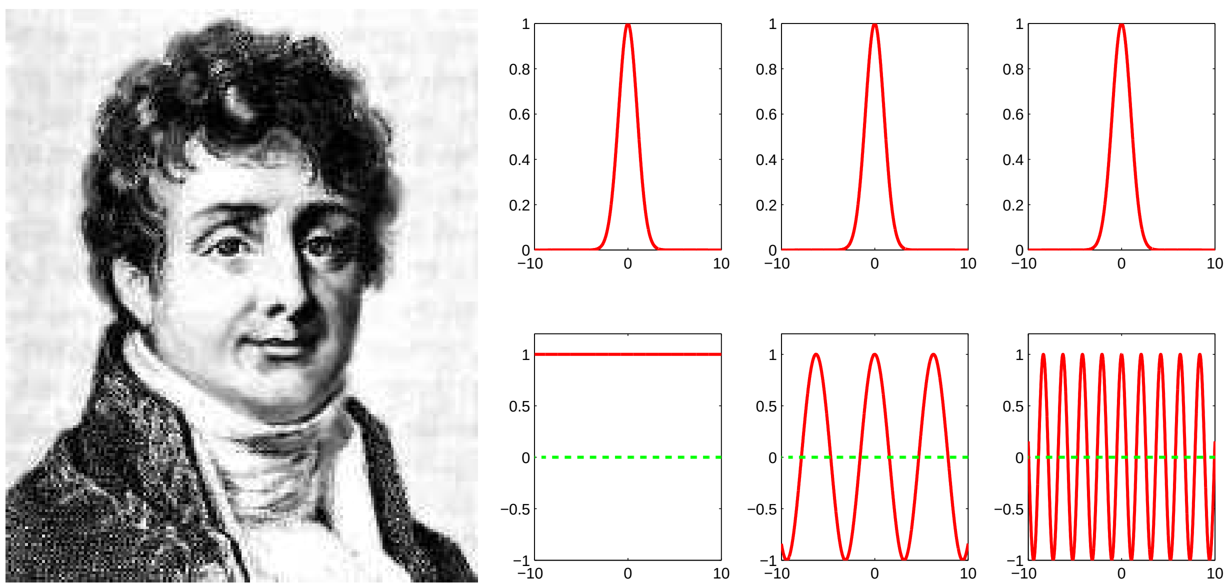
   - Reason: proportion is **scaling invariant**

$$X_j = \mu_j + \epsilon_j, \qquad \tilde{X}_j = \pm 3\mu_j + \epsilon_j, \qquad 1 \leq j \leq n$$

The Fourier Kingdom

- Function: nothing but superposition of waves
- A normal mixture is a mixture of waves

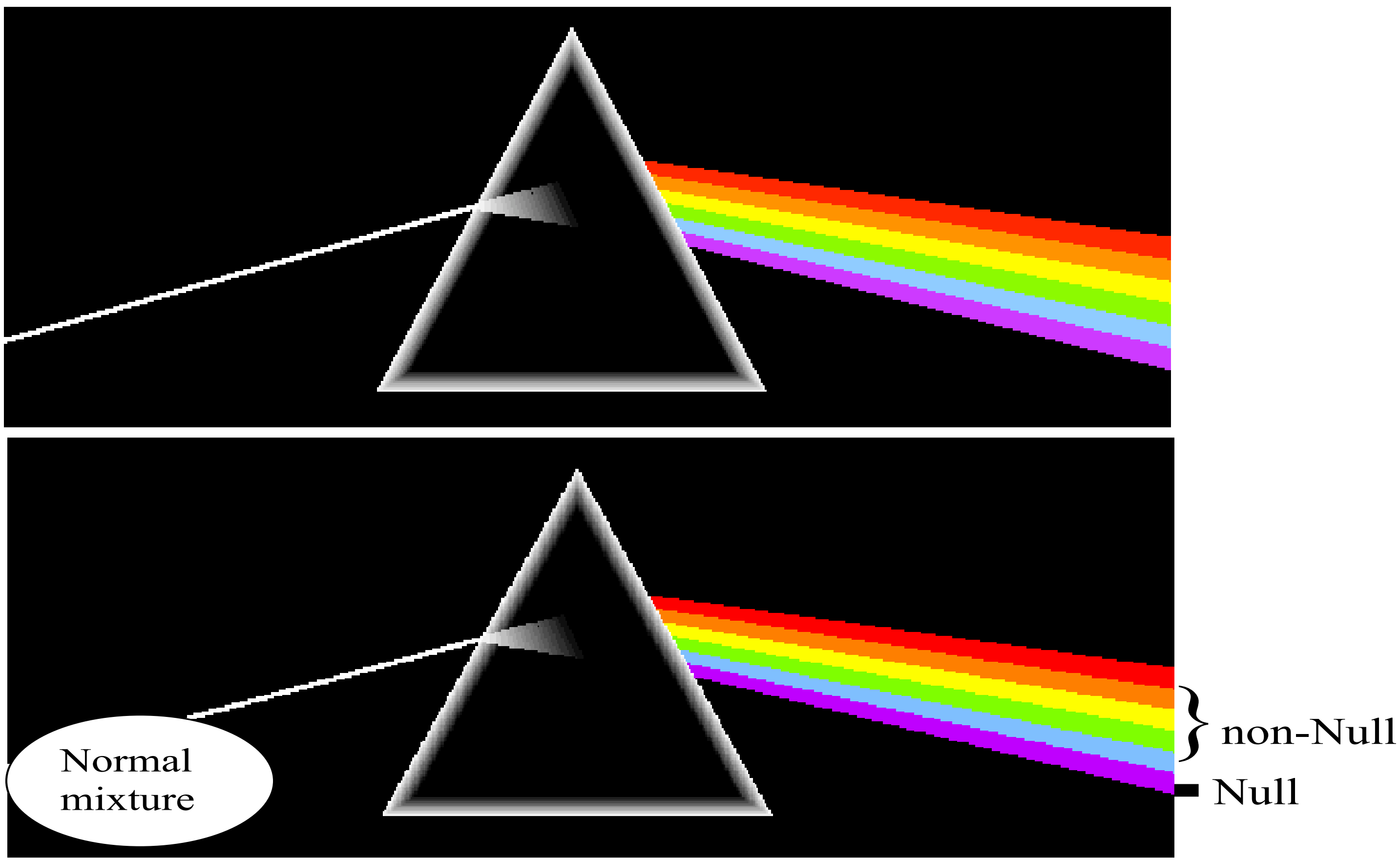
$$N(u,1) \quad \xrightarrow{FT} \quad e^{-\frac{t^2}{2}} \cdot e^{iut} \equiv \text{Amplitude} \cdot \text{Phase}$$



Left: Joseph Fourier (1768-1830). Right: $u = 0, 1, 3$

Reminiscent of Newton's Prism

Goal: Isolating the null component (and so an estimate of the proportion)



Phase Functions

Empirical phase function:

$$\varphi_n(t) = \varphi_n(t; X_1, \ldots, X_n) = \frac{1}{n} \sum_{j=1}^{n} e^{\frac{t^2}{2}} \cos(tX_j)$$

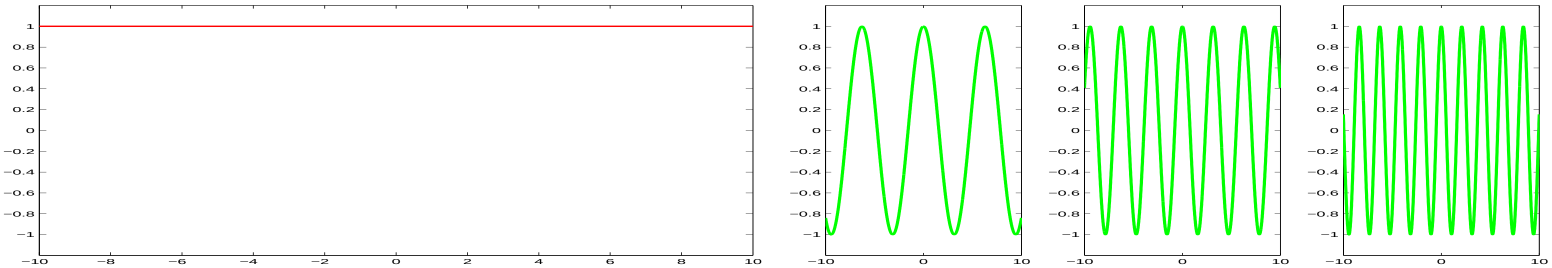
Underlying phase function:

$$\varphi(t) = \varphi(t; \mu, n) = \frac{1}{n} \sum_{j=1}^{n} \cos(t\mu_j)$$

Idea:

- Neglect stochastic fluctuations: $\varphi_n(t) \approx \varphi(t)$
- Use $\varphi(t)$ to construct an oracle estimator

Averaging: A Way to tell whether $\mu_j = 0$



$$\text{Phase}: \quad \varphi(t;\mu,n) = \frac{1}{n}\sum_{j=1}^{n}\cos(t\mu_j)$$

$$= [1-\epsilon_n(\mu)] + \frac{1}{n}\sum_{\{j:\mu_j\neq 0\}}\cos(t\mu_j)$$

Idea: average phase across a wide range of frequencies,

- the first term remains the same: $1-\epsilon_n(\mu)$
- the second term ≈ 0

Good Density: Choice of Weights

Call $\omega(\xi)$ *a good density* if

- a density function over $(-1, 1)$
- symmetric, continuous, and bounded
- $\omega(\xi) = g(1 - |\xi|)$; g: super-additive

Example: Triangle family with $\alpha \geq 1$

$$\omega(\xi) = \begin{cases} \frac{2}{\alpha+1}(1 - |\xi|)^{\alpha}, & |\xi| < 1 \\ 0, & \text{otherwise} \end{cases}$$

Universal Oracle Equivalence

Weighted empirical phase and underlying phase:

$$\psi_n(t;\omega) = \int_{-1}^{1} \omega(\xi)\varphi_n(t\xi)d\xi, \qquad \varphi_n(t) : \text{empirical phase}$$

$$\psi(t;\omega) = \int_{-1}^{1} \omega(\xi)\varphi(t\xi)d\xi, \qquad \varphi(t) : \text{underlying phase}$$

Theorem 1 (*Universal Oracle Equivalence*). If ω is good, then for any dimension n and normal means vector μ,

$$\epsilon_n(\mu) = \sup_t\{1 - \psi(t;\omega,\mu,n)\}$$

Interpretation

- Estimating the proportion reduces to estimating:

$$\sup_t \{1 - \psi(t;\omega)\} \equiv 1 - \lim_{t \to \infty} \psi(t;\omega)$$

- Where is the information?

Phase of high-frequency FT coefficients!

- Replacing $\psi(t;\omega)$ by $\psi_n(t;\omega)$ gives a *real estimator*:

$$1 - \psi_n(t;\omega) \approx 1 - \psi(t;\omega) \approx \epsilon_n(\mu)$$

- There is a trade-off in selecting t

Selecting t: Asymptotic Approach

- $t = \sqrt{2\gamma \log n}$: $\gamma \in (0, \frac{1}{2}]$
- $\psi_n(t;\omega)$: weighted empirical phase

Theorem 2 (*Uniform Consistency*). Suppose

1. (*summable*). $\frac{1}{n}\sum_{j=1}^{n} |\mu_j| \le r$
2. (*not very sparse*). true proportion $\ge n^{\gamma - 1/2}$
3. (*not very faint*). all nonzero means $\ge \frac{\log \log n}{\sqrt{\log n}}$

Then except an event with algebraically small prob.,

$$\lim_{n\to\infty}\left(\sup_{\{\mu \in \Theta_n(\gamma; r)\}} \left| \frac{[1-\psi_n(t;\omega)]}{\epsilon_n(\mu)} - 1 \right| \right) = 0$$

Selecting t: Adaptive Approach

1. Want: approaches adaptive for ω and small n
2. **Key**: the estimator $= \frac{1}{n}\sum_{j=1}^{n}[1 - g(X_j;\omega)]$; $g(X_j;\omega)$ has the largest 2^{nd} moment when $\mu_j = 0$
3. Adaptive selection of t:

$$t_n^*(\omega) = \max\{t : \left[s_0^2(t;\omega) + \frac{1}{n}\right] \leq \alpha_n\}$$

- α_n: specified tolerance for variance
- $s_0^2(t;\omega)$: variance when $\mu = 0$

Theorem 3 (*Adaptive Control on Variance*).

- For any n, ω, and normal means vector μ,

$$\mathrm{Var}(\psi(t_n^*(\omega);\omega)) \leq \alpha_n$$

- Theorem 2 continues to hold if $\alpha_n \to 0$ *slowly enough* (i.e. $t_n^*(\omega) \asymp \sqrt{\log n}$)

Advantage:

- $t_n^*(\omega)$ is non-stochastic, easy to calculate
- an adaptive control on variance (for n, ω, μ)

Extension to Heteroscedastic Gaussian Models

$$X_j \sim \begin{cases} N(0,1), & & H_j \text{ is true} \\ N(\mu_j, \sigma_j^2), & (\mu_j, \sigma_j) \neq (0,1), & \text{otherwise} \end{cases}$$

- found in many applications (e.g. microarray, CGH)
- handle many interesting situations
- proportion of non-null effects:

$$\epsilon_n = \frac{\#\{(\mu_j, \sigma_j) \neq (0,1)\}}{n}$$

Identifiability

1. Too broad to be identifiable: any density $\approx$ a Gaussian mixture (ℓ^1-metric)

2. Identifiable conditions

   - Elevated variances (Efron (2004))

$$\text{When } H_j \text{ is untrue :} \qquad \sigma_j \geq 1$$

   - Elevated means (CGH array):

$$\text{When } H_j \text{ is untrue :} \qquad \mu_j > 0$$

Main Results on Heteroscedastic Models

1. Elevated variances (Jin and Cai, JASA in press)
   - theoretic results successfully extended
   - applied to breast cancer microarray
2. Elevated means (Jin, Peng, and Wang, manuscript)
   - to accommodate heteroscedasticity, replace $i = \sqrt{-1}$ by $\sqrt{\sqrt{-1}}$

$$N(u, \sigma^2) \qquad \xrightarrow{FT} \qquad e^{-\frac{ut}{\sqrt{2}}} \cdot e^{i(\frac{\sigma^2}{2} + \frac{u}{\sqrt{2}} u)}$$

   - applied to lung cancer CGH array

- Empirical FT coefficients ≈ underlying FT coefficients
- In high-frequency underlying FT coefficients: null component sticks out

$$\left|\frac{1}{n}\sum_{j=1}^{n} e^{-\frac{\mu_j t}{\sqrt{2}}} e^{[i(\frac{\sigma_j^2 t^2}{2}+\frac{\mu_j t}{\sqrt{2}})]}\right|$$

$$=\left|e^{-\frac{it^2}{2}} \cdot \left\{[1-\epsilon_n] + \frac{1}{n}\sum_{\{j:\mu_j\neq 0\}} e^{-\frac{\mu_j t}{\sqrt{2}}} \cdot e^{i[\frac{(\sigma_j^2-1)t^2}{2}+\frac{\mu_j t}{\sqrt{2}}]}\right\}\right|$$

$$\approx[1-\epsilon_n]$$

Other Works on Estimating the Proportion

- Schweder (82), Storey (02), Genovese and Wasserman (04), Meinshausen and Rice (06)
- Langaas (05), Swanepoel (99)
- Only consistent under the *purity* condition

Purity Condition

- Introduced in Genovese and Wasserman (2004)
- $X_j \overset{iid}{\sim} (1-\epsilon_n) f_0 + \epsilon_n f$
  - f_0: $N(0,1)$, marginal density of null effects
  - f: marginal density of non-null effects
- Purity condition

$$\inf_{-\infty < x < \infty} \left\{ \frac{f(x)}{f_0(x)} \right\} = 0$$

- For the sake of *identifiability*

Comparisons

Meinshausen and Rice type approaches

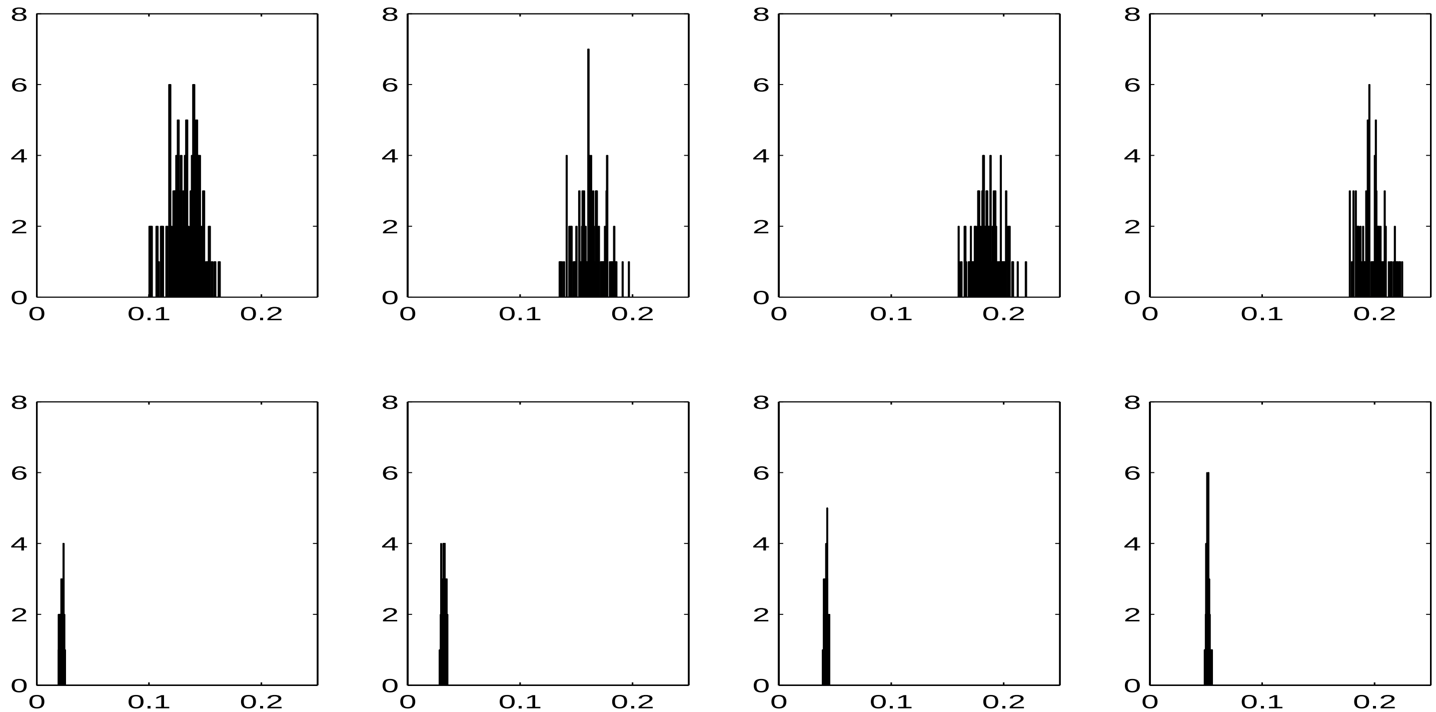
- Use fully non-parametric models
- Consistent only in the *purity* regime

Our approaches:

- Use Gaussian models
- Remove the hurdle of identifiability
- Successful beyond the *purity* regime

Comparisons Using Simulated Data

- Fix $n = 10^5$ and $\epsilon_n = 0.2$
- Pick $a = \frac{1}{4} \times 2, 3, 4, 5$
- Conduct 100 simulation cycles
  1. Generate $n(1 - \epsilon_n)$ null effects $X_j \sim N(0, 1)$
  2. Generate $n\epsilon_n$ non-null effects $X_j \sim N(\mu_j, 1)$
     - $|\mu_j| \overset{iid}{\sim} \text{Uniform}(a, a + 1)$
     - $\text{sgn}(\mu_j) \overset{iid}{\sim} \{-1, 1\}$
  3. Implement our adaptive approach with $\alpha = 0.015$ (ω: triangle density)
  4. Implement Meinshausen and Rice's approach

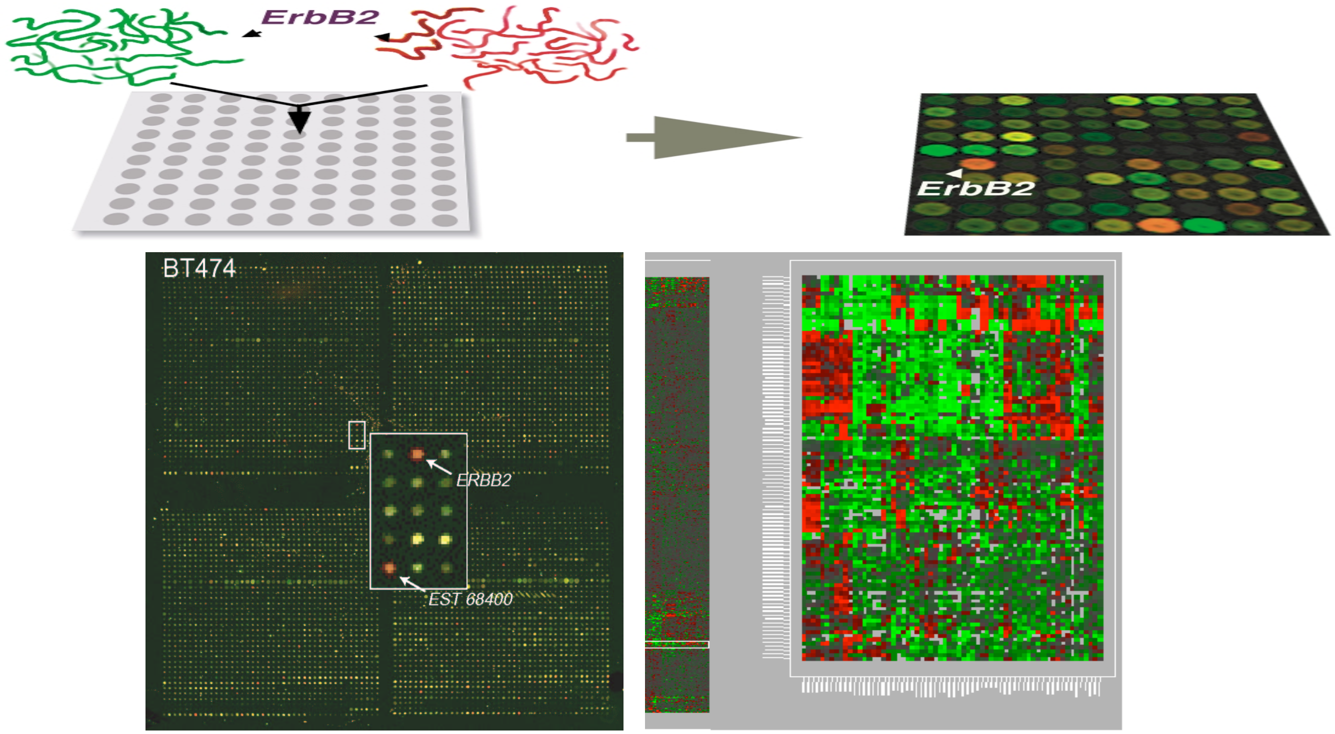


- $n = 10^5$, $\epsilon_n = 0.2$, $a = \frac{1}{4} \times 2, 3, 4, 5$
- Each cycle: $\quad n(1-\epsilon_n)$ null effects $X_j \sim N(0,1)$

  $n\epsilon_n$ non-null effects $X_j \sim N(\mu_j, 1)$

  $$|\mu_j| \overset{iid}{\sim} \text{Uniform}(a, a+1), \quad \text{sgn}(\mu_j) \overset{iid}{\sim} \{-1, 1\}$$

Applications to Lung Cancer CGH array



Paired CGH profile (left) and mRNA profile (right)

CGH: Comparative Genomic Hybridization.

Abstraction

- 23 tumor cells, same set of 25736 genes
- N_j: log-intensity of CGH profile, measures DNA copy number alternation; j: j-th gene
- R_j: log-intensity of mRNA profile, measures RNA expression level

Interested in:

- proportion of genes where N_j and R_j are correlated
- conjectured to be large (2/3)

Amplification and Deletion

Terminology:

- N_j: DNA copy number alternation of j-th gene
- Call DNA copy number *amplification*, *deletion*, or *no alternation* if $N_j > 0$, < 0, or $= 0$

Accordingly, genes split into two groups:

- *Amplification (13356 genes).* ≥ 1 amplifications across 23 cells
- *Deletion (11283 genes).* ≥ 1 deletion across 23 cells

Multiple Testing Setup (Amplification)

- n hypotheses:

$$H_j : \qquad N_j \text{ and } R_j \text{ not correlated}$$

  equivalent to (roughly)

$$H_j : \qquad (R_j | N_j > 0) \;\; =_d \;\; (R_j | N_j = 0)$$

- n test statistics:

$$X_j = \bar{\Phi}^{-1}(p_j) \qquad p_j\text{:} \quad \text{p-value based on rank test}$$

- (*Elevated mean*). If H_j is true: $X_j \sim N(0,1)$
  Otherwise: $\quad X_j \sim N(\mu_j, \sigma_j^2), \quad \mu_j > 0$

	Jin, Peng, Wang	MR	Efron/GR/Storey
Amplificatioin	0.585	0.466	0.406
Deletion	0.539	0.464	0.418

Take Home Messages

1. Constructed universal oracle equivalence of the proportion
2. Developed estimators which are uniformly consistent over a wide class of parameters.
3. Applied to lung cancer CGH array

Acknowledgements

Javier Cabrera	Rutgers University
David Donoho	Stanford University
Bradley Efron	Stanford University
Jianqing Fan	Princeton University
Chris Genovese	Carneige Mellon University
MarK Low	University of Pennsylvania
Eric Kolaczyk	Boston University
Herman Rubin	Purdue University
Jeffe Scargle	NASA Ames Research Center
Paul Shaman	University of Pennsylvania
Larry Wasserman	Carneige Mellon University
Cun-Hui Zhang	Rutgers University