UNC, Stat & OR

High Dimension Low Sample Size Asymptotics

J. S. Marron
School of Data Science and Society
Dept. of Statistics and Operations Research,
University of North Carolina

July 21, 2024

UNC, Stat & OR

HDLSS Asymptotics

High Dimension Low Sample Size

Terminology Coined in

Hall, Marron & Neeman (2005)

UNC, Stat & OR

Motivation of HDLSS Asy's

Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Gener'n Sequen'g)
- Deep look at "gene components"

- Gene studied here: CDKN2A
- Goal: *Study Alternate Splicing*
- Sample Size, $n = 180$
- Dimension, $d \sim 1700$

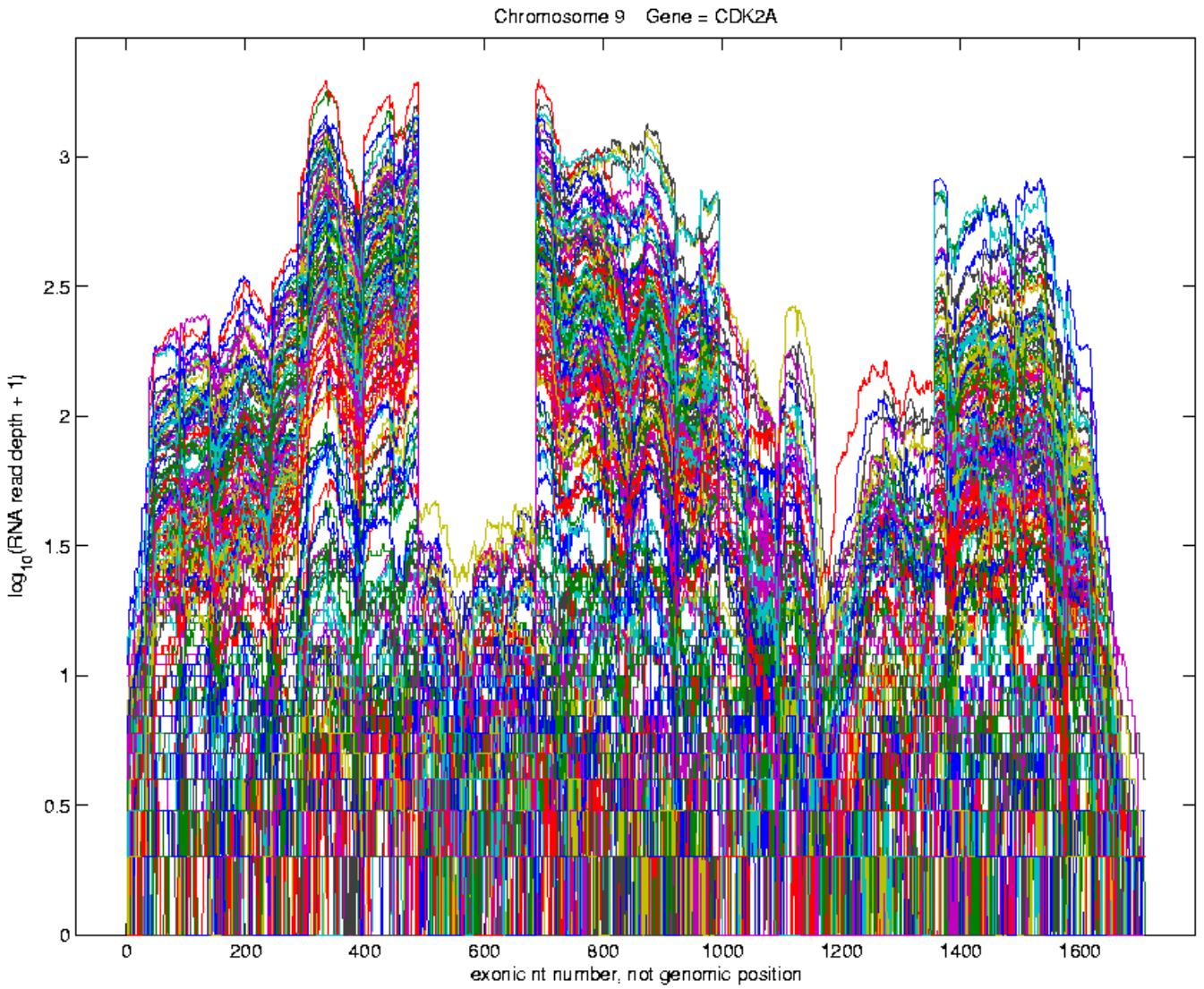
UNC, Stat & OR

Motivation of HDLSS Asy's

Simple 1st View: Curve Overlay

(log scale)

Thanks to Matt Wilkerson

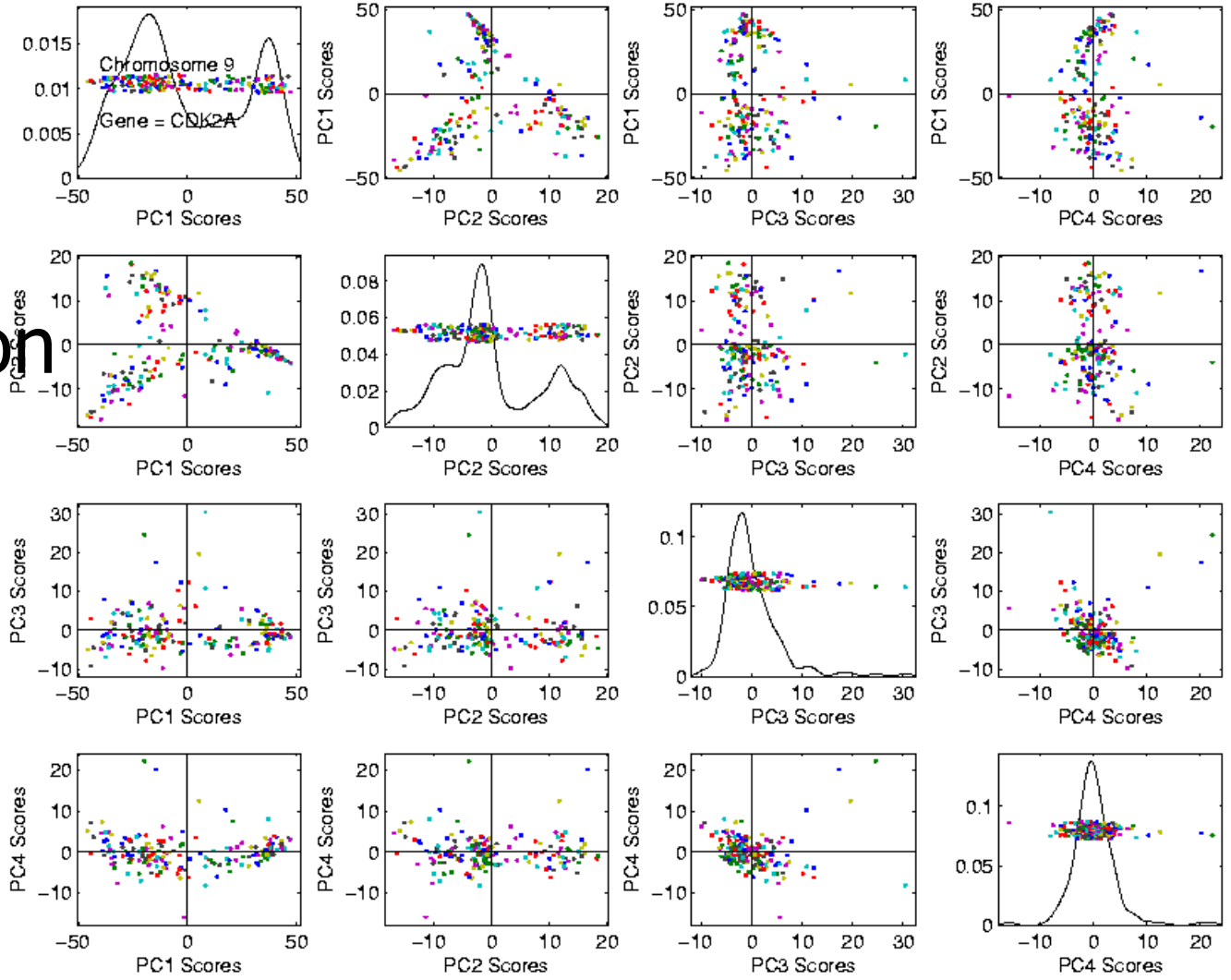


UNC, Stat & OR

Motivation of HDLSS Asy's

Often Useful Population View:

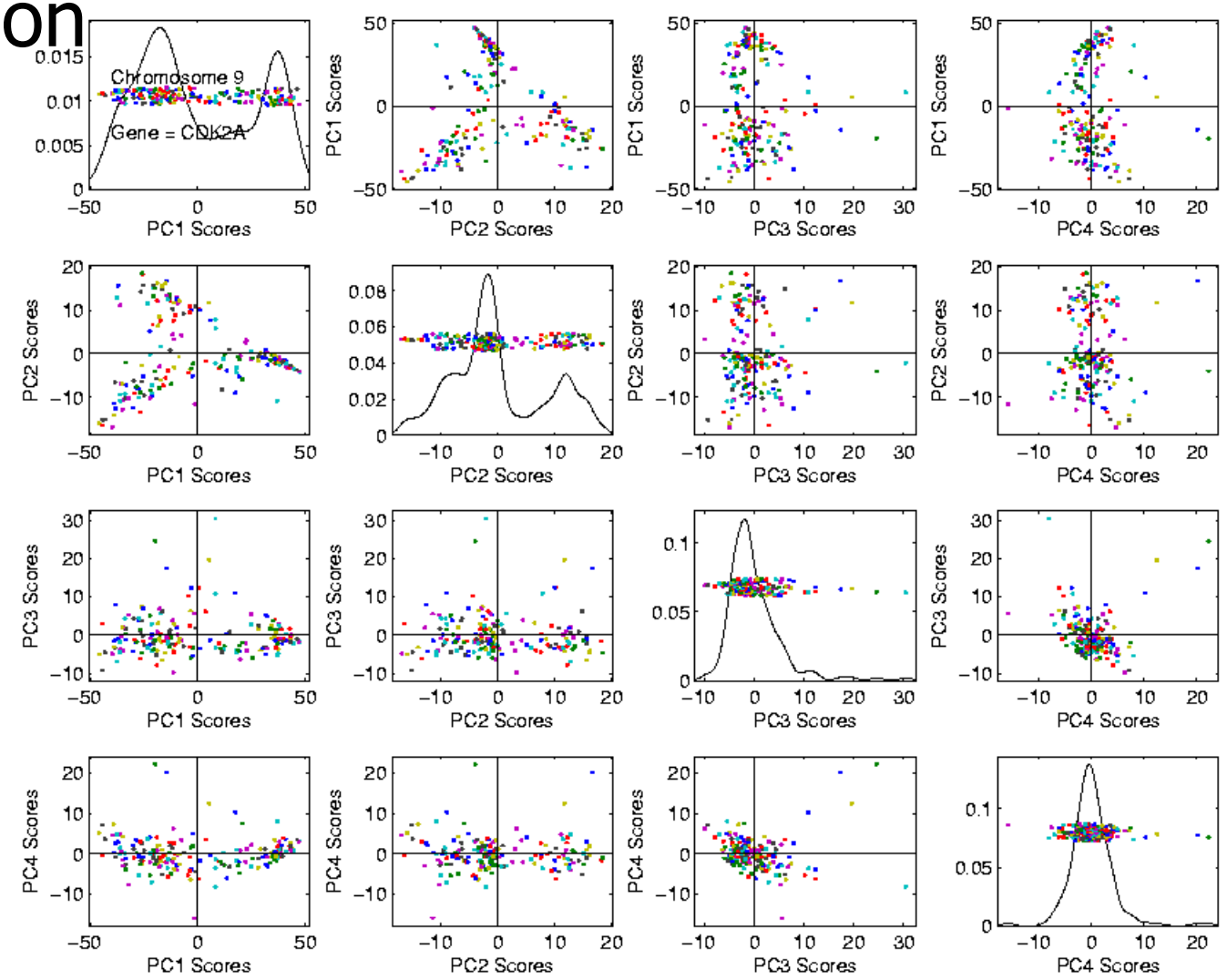
PCA Scores



UNC, Stat & OR

Motivation of HDLSS Asy's

Suggestion Of Clusters ???

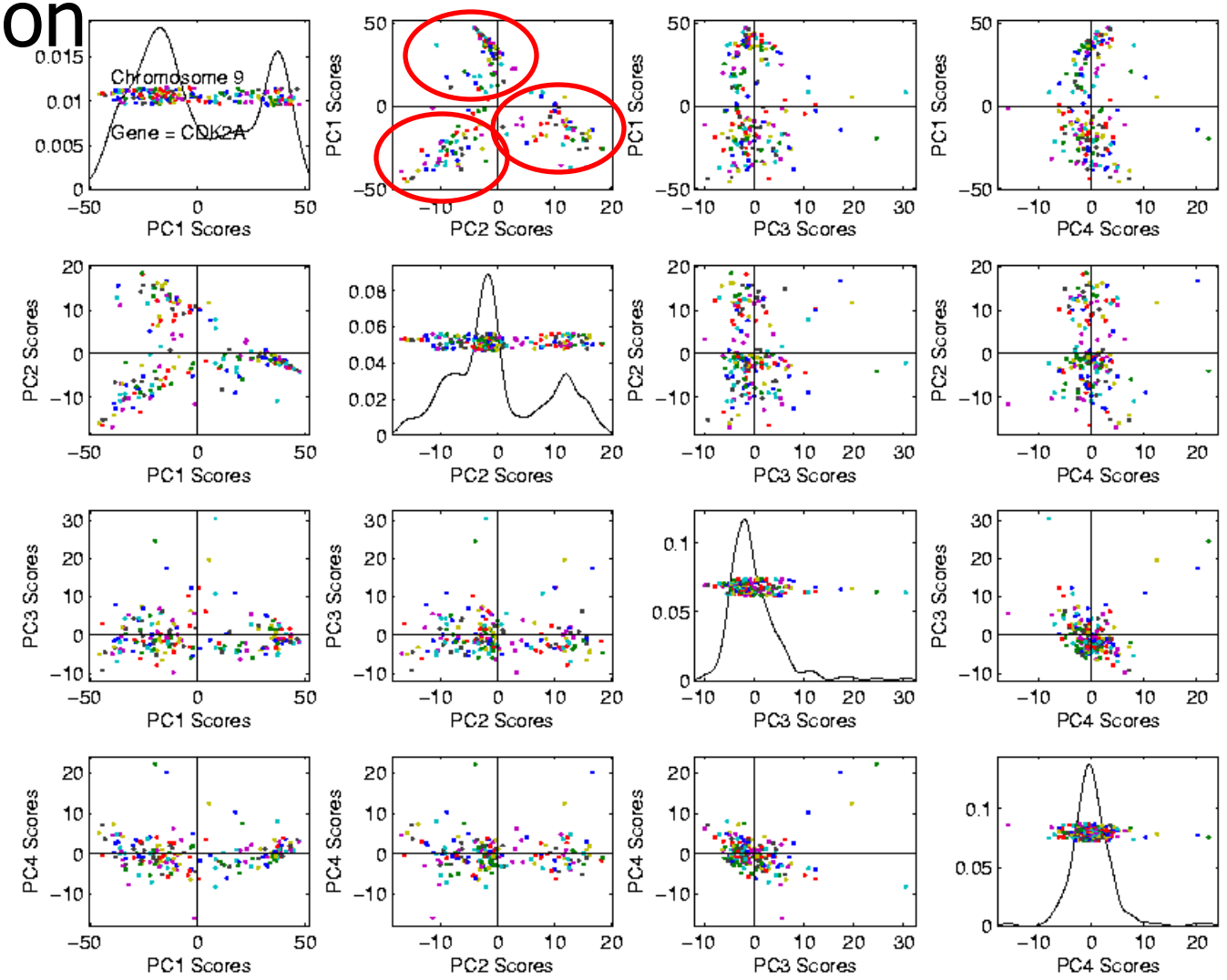


UNC, Stat & OR

Motivation of HDLSS Asy's

Suggestion Of Clusters

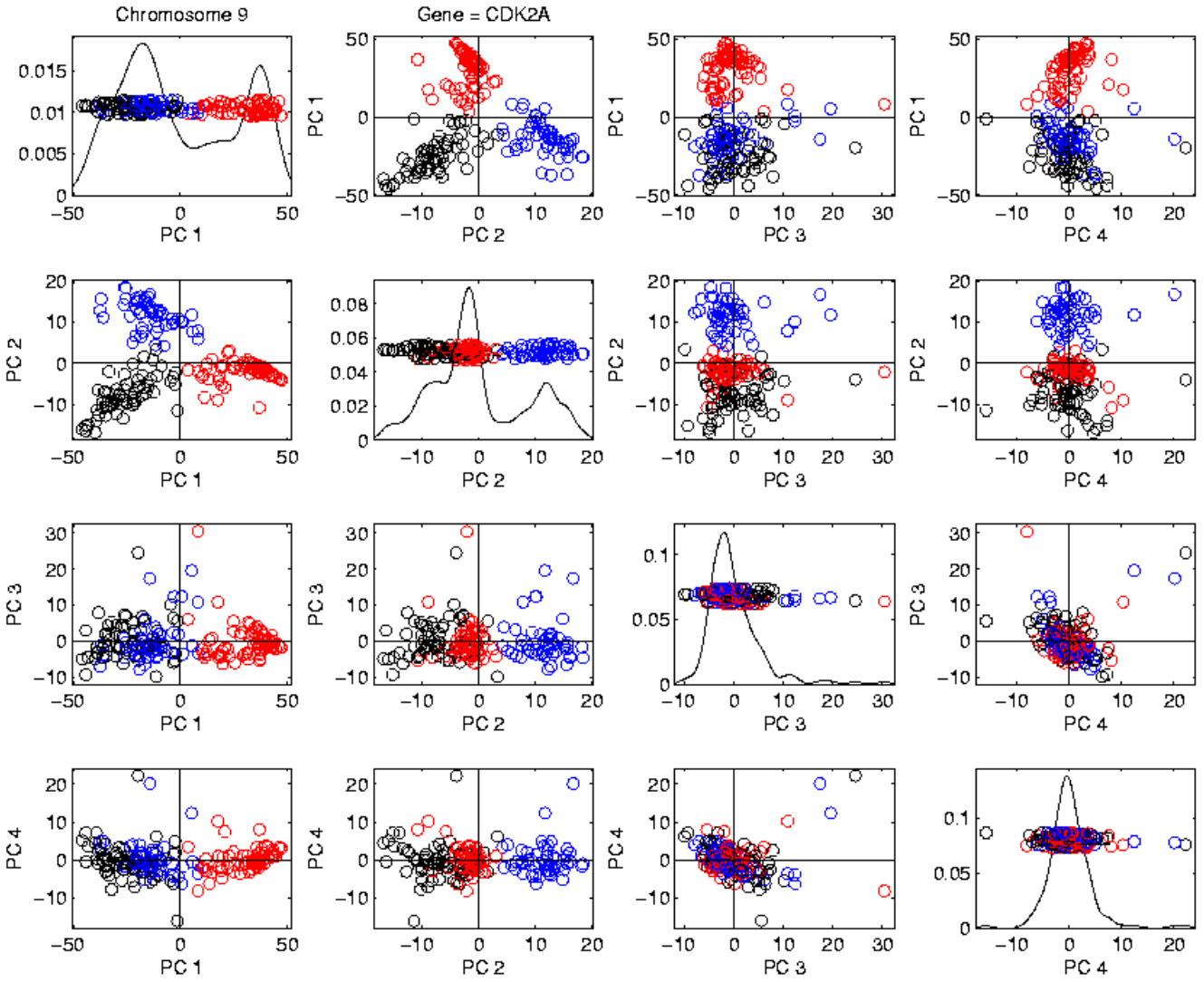
Which Are These?



UNC, Stat & OR

Motivation of HDLSS Asy's

Manually Brush Clusters

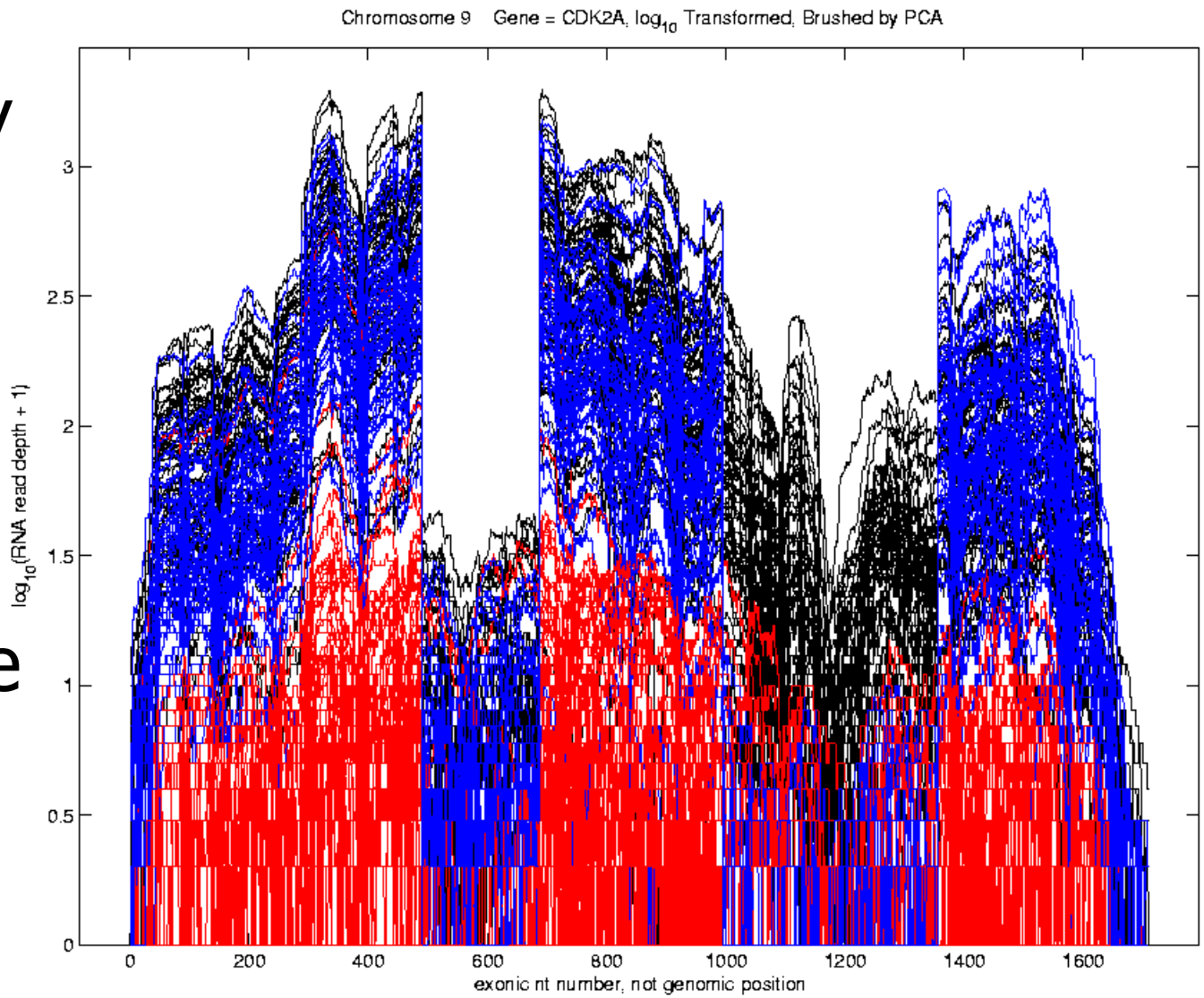


UNC, Stat & OR

Motivation of HDLSS Asy's

Manually Brush Clusters

Clear Alternate Splicing



UNC, Stat & OR

Motivation of HDLSS Asy's

Consequences of this Visualization:

- ✓ Lead to Full Genome Screening Method
  SigFuge
- ✓ Important Component: SigClust
  (Which Clusters are *Really There*?)
- ✓ Found New Splices
  (Now Been Biologically Verified)

UNC, Stat & OR

Motivation of HDLSS Asy's

Important Points

- ✓ PCA found *Important Structure*
- ✓ In High Dimensional Data Analysis

UNC, Stat & OR

HDLSS Asymptotics

Modern Mathematical Statistics:

- Based on *asymptotic* analysis
- I.e. Uses limiting operations
- Very often $\lim_{n \to \infty}$

Workhorse Method for Much Insight:

- ❖ Laws of Large Numbers (Consistency)
- ❖ Central Limit Theorems

  (Quantify Errors, Basis of Inference)

UNC, Stat & OR

HDLSS Asymptotics

Modern Mathematical Statistics:

- Based on *asymptotic* analysis
- I.e. Uses limiting operations
- Very often $\lim_{n \to \infty}$
- Occasional misconceptions:
  - Indicates behavior for *large samples*
  - Thus only makes sense for "large" samples
  - Models phenomenon of "increasing data"
  - So other flavors are useless???

Sometimes Ask Junior Researchers Why They Do Asymptotics

UNC, Stat & OR

HDLSS Asymptotics

Modern Mathematical Statistics:

- Based on *asymptotic* analysis
- Real Reasons:
  - Approximation provides insights
  - Can find simple underlying structure
  - In complex situations
- Thus various flavors are fine:

$$\lim_{n\to\infty} \qquad \lim_{d\to\infty} \qquad \lim_{n,d\to\infty} \qquad \lim_{\sigma\to 0}$$

Even desirable! (find additional insights)

UNC, Stat & OR

HDLSS Asymptotics

Which asymptotics?

All Interesting "Blessings"

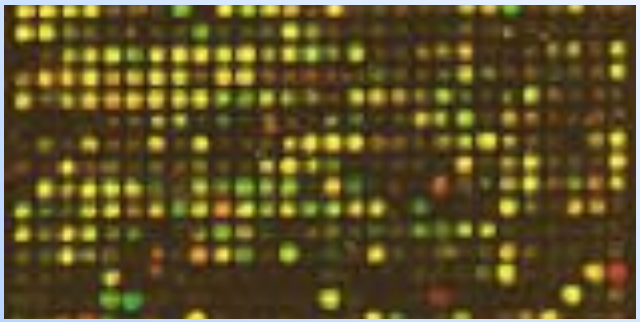
- $n \to \infty$ & $d \to \infty$
  - $n \gg d$: close to "classical" (Portnoy)
  - $n \sim d$: random matrices (Johnstone)
  - $d \gg n$: "ultra high dimension" (Fan)?
- HDLSS asymptotics: n fixed, $d \to \infty$

UNC, Stat & OR

HDLSS Asymptotics

HDLSS asymptotics: n fixed, $d \to \infty$

- Follow typical "sampling process"?
  - Microarrays: # genes bounded
  - Proteomics, SNPs, …

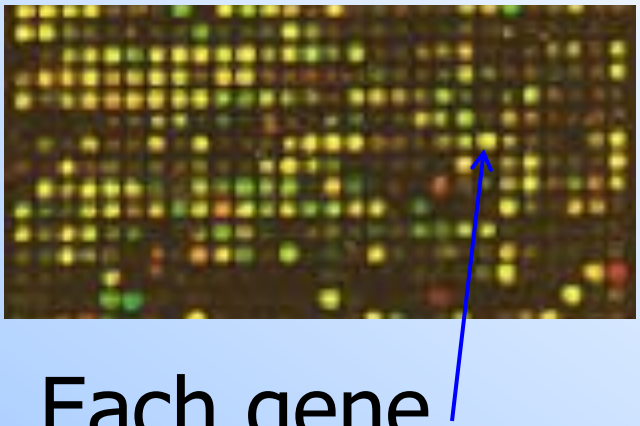


UNC, Stat & OR

HDLSS Asymptotics

HDLSS asymptotics: n fixed, $d \to \infty$

- Follow typical "sampling process"?
  - Microarrays: # genes bounded
  - Proteomics, SNPs, …

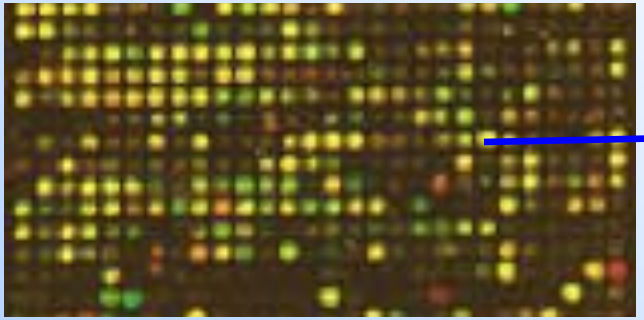


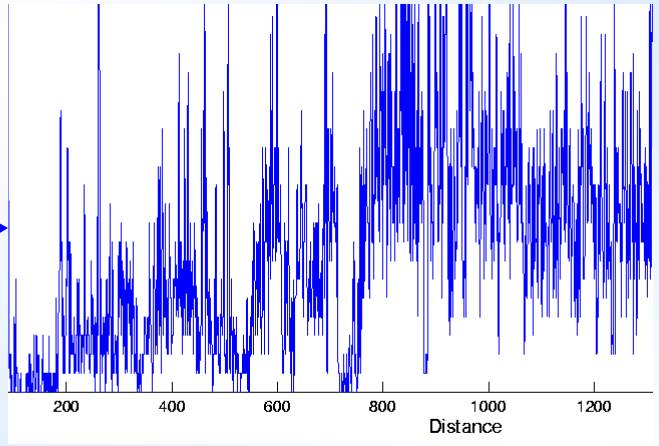
UNC, Stat & OR

HDLSS Asymptotics

HDLSS asymptotics: n fixed, $d \to \infty$

- Follow typical "sampling process"?
  - Microarrays: # genes bounded
  - Proteomics, SNPs, …





Next Gen Sequencing

Now called "RNA-Seq"

UNC, Stat & OR

HDLSS Asymptotics

HDLSS asymptotics: n fixed, $d \to \infty$

- Follow typical "sampling process"?
  - Microarrays: # genes bounded
  - Proteomics, SNPs, …
- A moot point, from perspective:

Asymptotics are a tool for finding *simple structure* underlying complex entities

UNC, Stat & OR

HDLSS Asymptotics

HDLSS asymptotics: n fixed, $d \to \infty$

- Say anything, as noise level increases???

Yes, there exists simple, perhaps surprising, underlying structure

HDLSS Asymptotics

Personal Observations:

HDLSS world is…

- Surprising (many times!)

  [Think I've got it, and then …]

- Mathematically Beautiful (?)
- Practically Relevant
- Publishable???

Key Point:
Must do the Math

UNC, Stat & OR

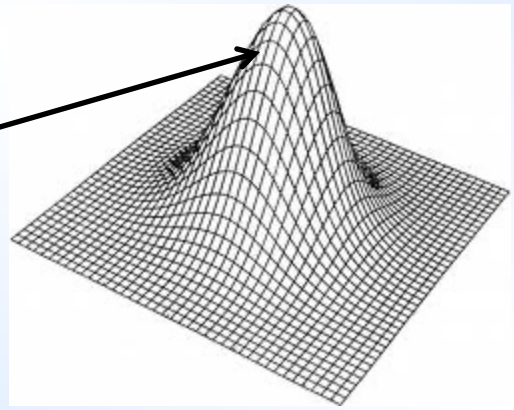
HDLSS Asymptotics: Simple Paradoxes

For d dimensional *Standard Normal* dist'n:

$$\tilde{\boldsymbol{z}} = \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_d \end{pmatrix} \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$$

Where are the Data?

Near Peak of Density?



UNC, Stat & OR

HDLSS Asymptotics: Simple Paradoxes, I

For d dim'al "Standard Normal" dist'n:

$$\underline{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} \sim N_d\left(\underline{0}, I_d\right)$$

Euclidean Distance to Origin (as $d \to \infty$):

$$\left\|\underline{Z}\right\| = \sqrt{d} + O_p(1)$$

- Data lie roughly on surface of sphere of radius $\sqrt{d}$
- Yet origin is point of "highest density"???
- Paradox resolved by:

"density w. r. t. Lebesgue Measure"

UNC, Stat & OR

HDLSS Asymptotics: Simple Paradoxes, I

- Paradox resolved by:

density w. r. t. Lebesgue Measure

- Consider *Volume of Unit Sphere* in $\mathbb{R}^d$
- Find As: Integral In Sph'l Coordinates

$$V_d = \iiint J d\theta_1 \cdots d\theta_d dr$$

HDLSS Asymptotics: Simple Paradoxes, I

- Paradox resolved by:

density w. r. t. Lebesgue Measure

- *Consider Volume of Unit Sphere in* $\mathbb{R}^d$
- Find As: Integral In Sph'l Coordinates

$$V_d = \underbrace{\iiint}_{} J d\theta_1 \cdots d\theta_d dr$$

- Look At Integrand w.r.t. r
- Can Show: Puts ~ All Weight Near $r = 1$

UNC, Stat & OR

HDLSS Asymptotics: Simple Paradoxes, I

- Paradox resolved by:

density w. r. t. Lebesgue Measure

- ✓ Lebesgue Measure Pushes Mass Out
- ✓ Density Pulls Data In
- ✓ $\sqrt{d}$ Is The Balance Point

UNC, Stat & OR

HDLSS Asymptotics: Simple Paradoxes, I

As $d \to \infty$, $\quad \|\tilde{\mathbf{z}}\| = \sqrt{d} + O_p(1)$

Important Philosophical Consequence:

$\nexists$ "Average People"

Parents Lament:

Why Can't I Have Average Children?

Theorem: Impossible (over many factors)!

UNC, Stat & OR

HDLSS Asymptotics: Simple Paradoxes, II

For d dim'al "Standard Normal" dist'n:

$$\underline{Z}_1 \text{ indep. of } \underline{Z}_2 \sim N_d\left(\underline{0}, I_d\right)$$

Euclidean Dist. between $\underline{Z}_1$ and $\underline{Z}_2$ (as $d \to \infty$):

Distance tends to *non-random* constant:

$$\left\|\underline{Z}_1 - \underline{Z}_2\right\| = \sqrt{2d} + O_p(1)$$

Can extend to $Z_1, \cdots, Z_n$

Where do they all go???

(We can only perceive 3 dimensions)

UNC, Stat & OR

HDLSS Asymptotics: Simple Paradoxes, I

Reason For This

- Perceptual System from Ancestors
- They Needed to Find Food
- Food Exists in 3-d World

(We can only perceive 3 dimensions)

UNC, Stat & OR

HDLSS Asymptotics: Simple Paradoxes, III

For d dim'al "Standard Normal" dist'n:

$$\underline{Z}_1 \text{ indep. of } \underline{Z}_2 \sim N_d\left(\underline{0}, I_d\right)$$

High dim'al Angles (as $d \to \infty$):

$$Angle\left(\underline{Z}_1, \underline{Z}_2\right) = 90^\circ + O_p\left(d^{-1/2}\right)$$

- "Everything is orthogonal"???
- Where do they all go???

  (again our perceptual limitations)
- Key Point: 1st order structure is non-random

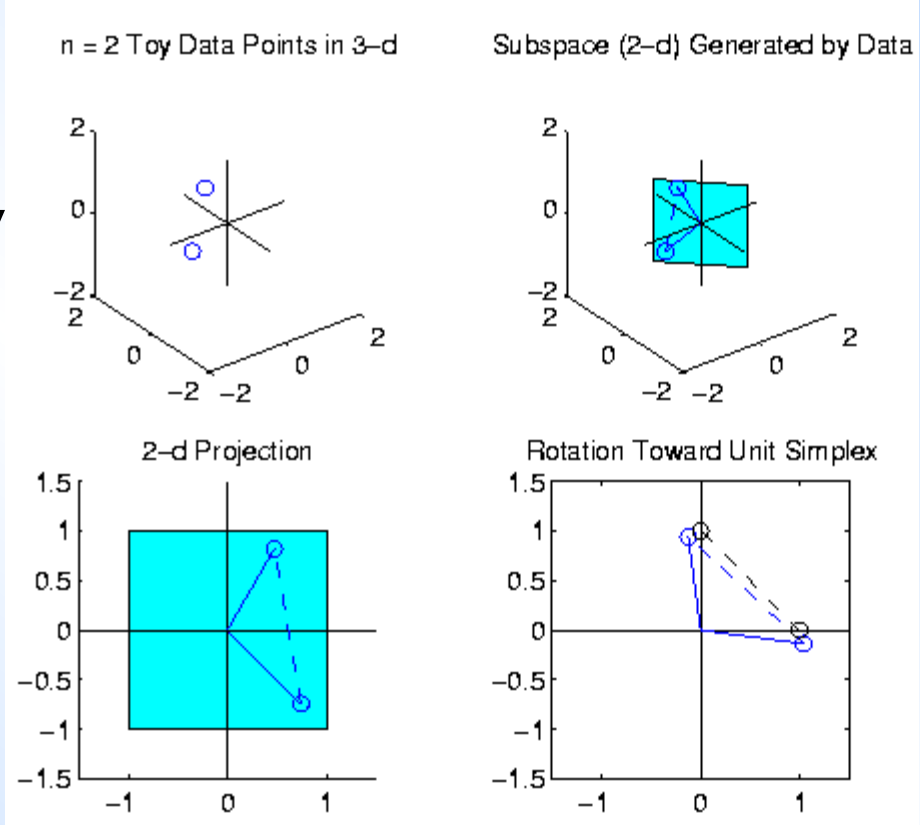
HDLSS Asy's: Geometrical Representation, I

Assume $\underline{Z}_1, \cdots, \underline{Z}_n \sim N_d(\underline{0}, I_d)$, let $d \to \infty$

Study Subspace Generated by Data

- Hyperplane through 0, of dim'n n
- Points are "nearly equidistant to 0", & dist $\sqrt{d}$
- Within plane, can "rotate towards $\sqrt{d}\times$ Unit Simplex"
- *All Gaussian data sets* are "near Unit Simplex Vertices"!!!

"Randomness" *appears only in rotation* of simplex



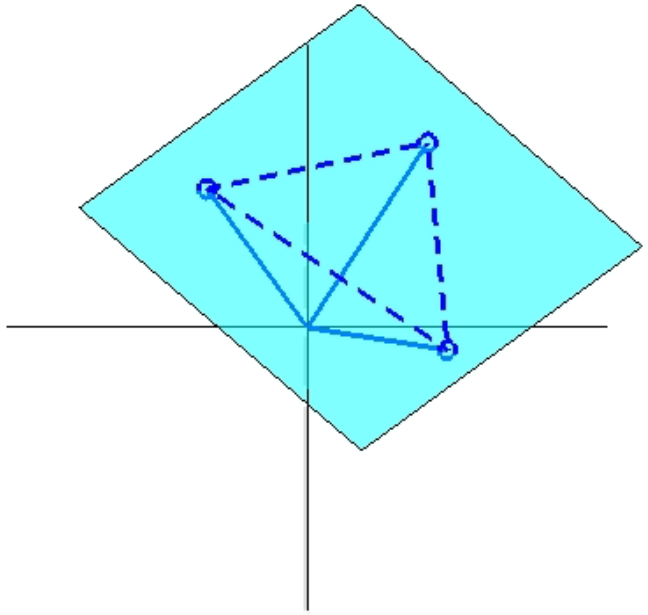
With P. Hall & A. Neeman

HDLSS Asy's: Geometrical Representation, II

Assume $\underline{Z}_1, \cdots, \underline{Z}_n \sim N_d(\underline{0}, I_d)$, let $d \to \infty$

Study Hyperplane Generated by Data

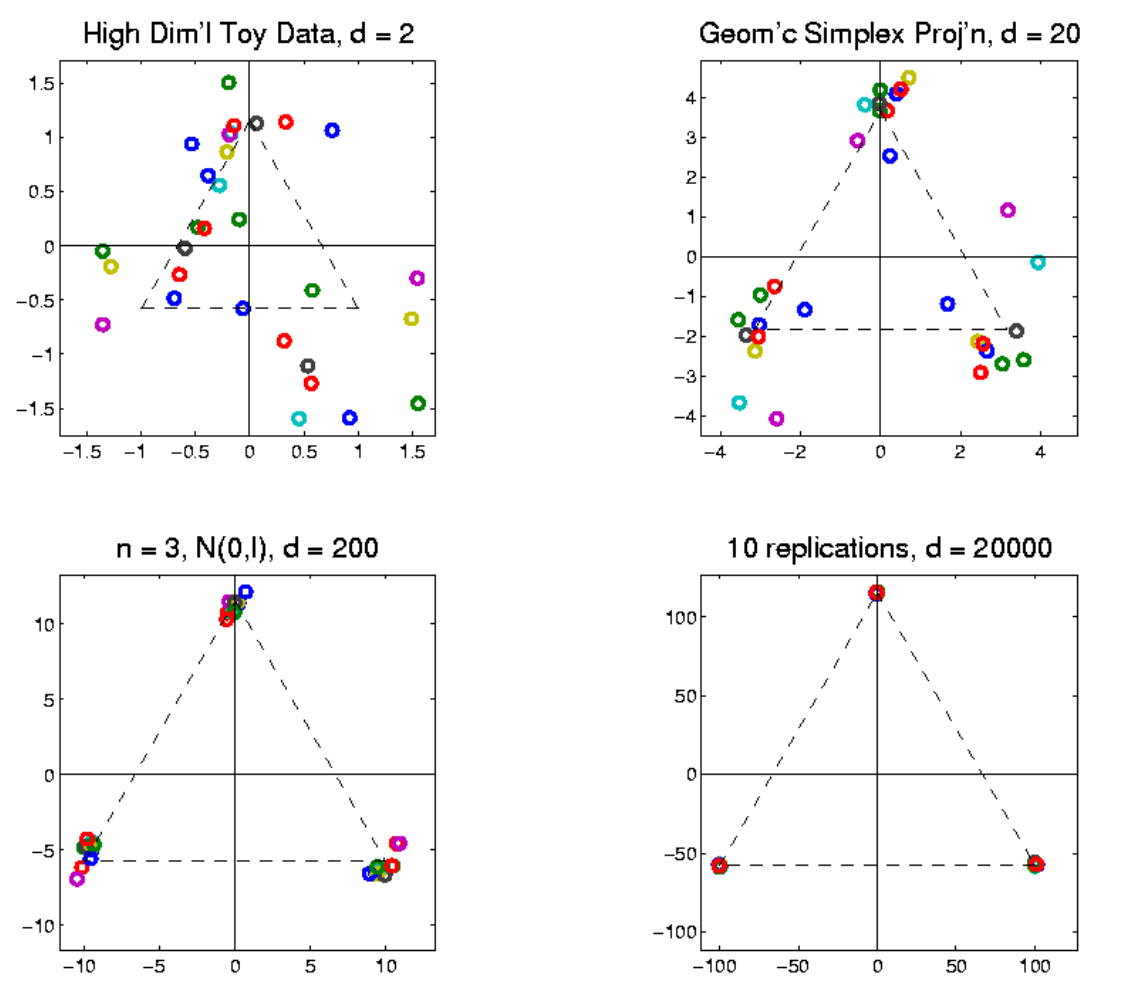
- $n-1$ dimensional hyperplane
- Points are pairwise equidistant, dist $\sim\sqrt{2d}$
- Points lie at vertices of "$\sqrt{2d}\times$ regular $n-$ hedron"
- Again "randomness in data" is *only in rotation*
- Surprisingly rigid structure in data?



Key Point:
Must do the Math

HDLSS Asy's: Geometrical Representation, III

Simulation View: shows "rigidity after rotation"



HDLSS Asy's: History & Assumptions

Hall, Marron and Neeman (2005)

- Above $\tilde{\boldsymbol{x}} \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$ is too strong
- So Assumed ρ *Mixing Condition*

Common in Time Series

Far Apart Entries Uncorrelated

In $\lim_{d \to \infty}$

$$\begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_i \\ \vdots \\ \tilde{x}_d \end{pmatrix}$$

UNC, Stat & OR

HDLSS Asy's: History & Assumptions

Hall, Marron and Neeman (2005)

Publication History:

Rejected by *Biometrika*

"ρ-Mixing Along Vector Not Practical"

Hence Published in *JRSS-B*

HDLSS Asy's: History & Assumptions

Hall, Marron and Neeman (2005)

Later Realization:

This Mixing is Very Natural in

Genome Wide Association Studies

1st such: Klein et al (2005)

UNC, Stat & OR

HDLSS Asy's: History & Assumptions

Genome Wide Association Study

Data Objects: Vectors of Genetic Variants, at known chromosome locations
(Called SNPs)

Discrete (takes on 2 or 3 values)

Dimension d as large as ~5 million
(can be reduced, e.g. $d \sim 20000$)

UNC, Stat & OR

HDLSS Asy's: History & Assumptions

Genome Wide Association Study

Data Objects: Vectors of Genetic Variants, at known chromosome locations

(Called SNPs)

Sexual Reproduction $\Rightarrow$ ρ - Mixing

Actually Very Natural "In Practice"

UNC, Stat & OR

HDLSS Asy's: History & Assumptions

View of GWAS data: Pairwise Angles

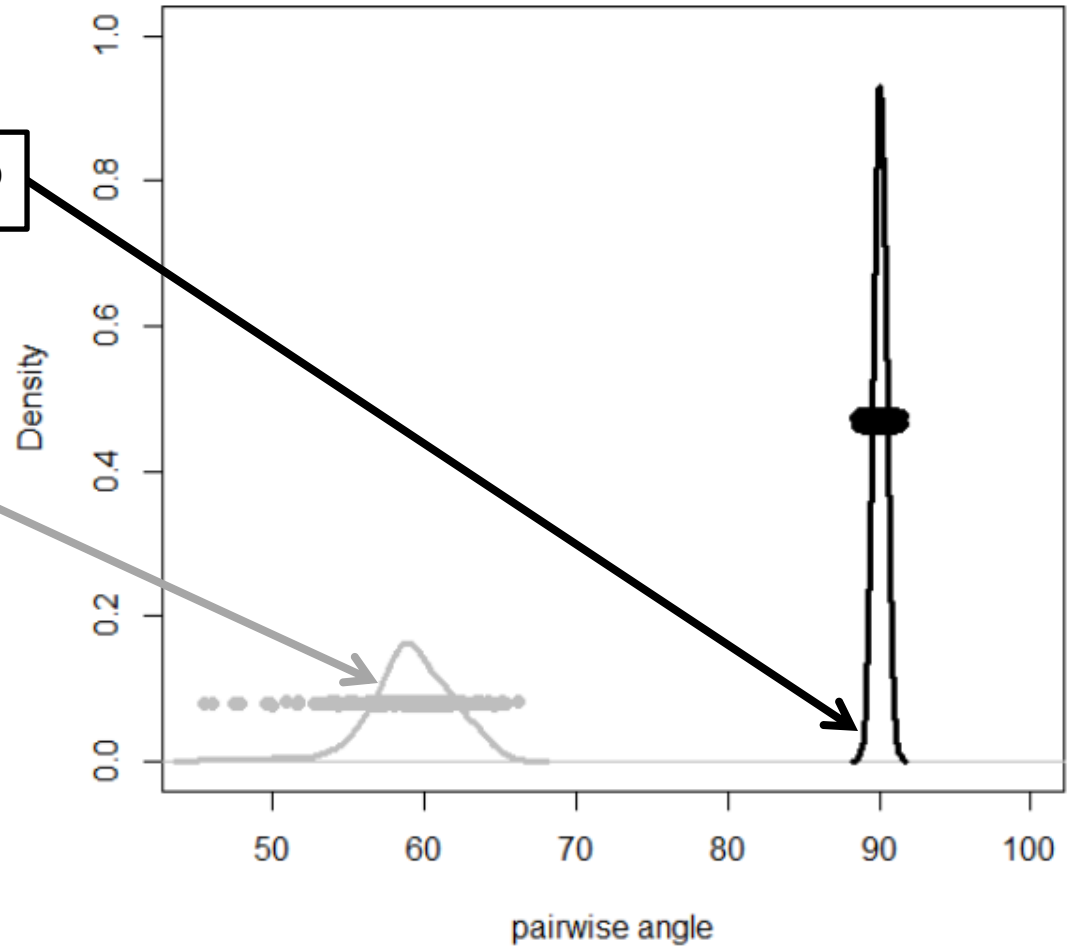
d = 20,000

Most Around 90°

Others Are Related

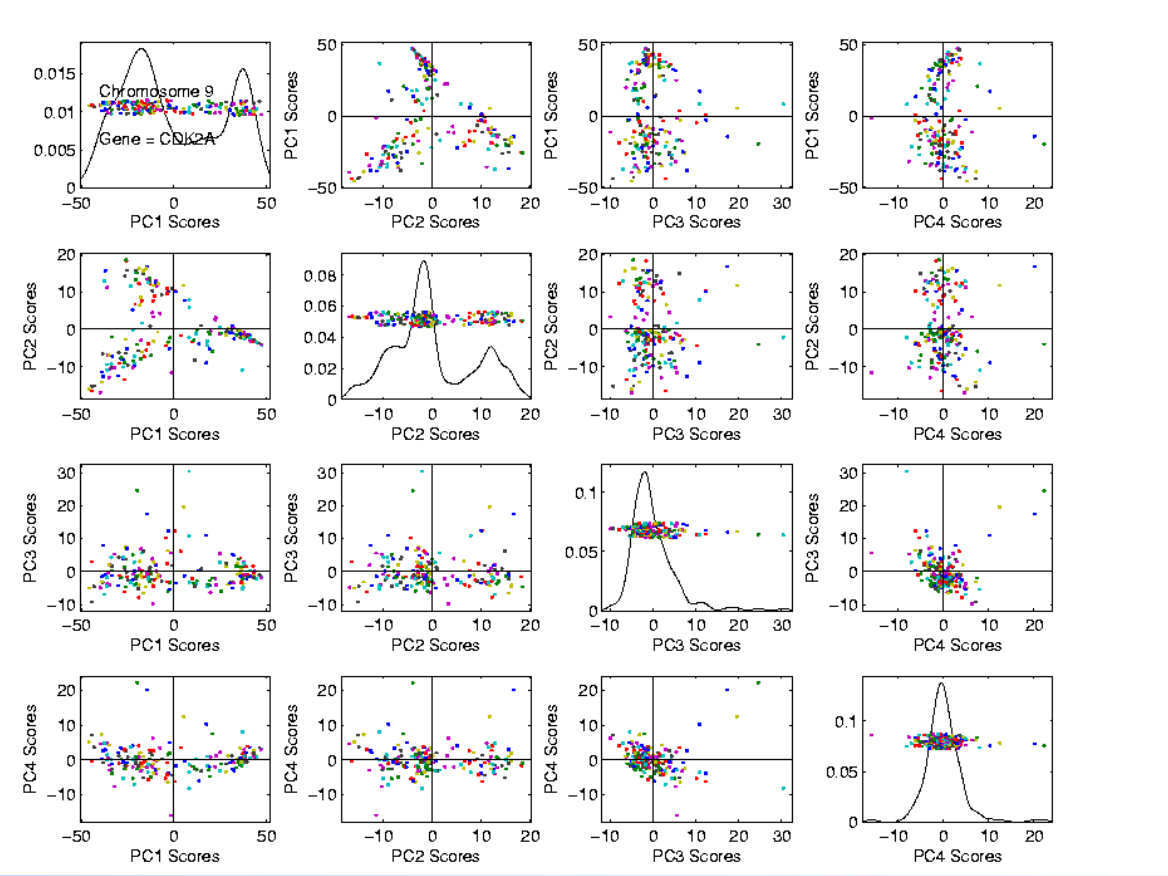
Note: Discrete, NOT Gaussian

Thanks to Yihui Zhou



HDLSS Math. Stat. of PCA

Next Study Principal Component Analysis



UNC, Stat & OR

HDLSS Math. Stat. of PCA

Next Study Principal Component Analysis

For Centered $d \times n$ Data Matrix, SVD is

$$\breve{\boldsymbol{X}} = \breve{\boldsymbol{U}}\breve{\boldsymbol{S}}\breve{\boldsymbol{V}}^t = \sum_{k=1}^{r} \boldsymbol{u}_k s_k \boldsymbol{v}_k^t$$

Loadings (Directions) Vectors

Scores (Projection Coefficients)

Equivalent to Eigen-Analysis of Cov. Matrix

UNC, Stat & OR

HDLSS Math. Stat. of PCA

Consistency & Strong Inconsistency:

Spike Covariance Model (Johnstone & Paul)

For Eigenvalues: $\lambda_{1,d} = d^{\alpha}, \lambda_{2,d} = 1, \cdots, \lambda_{d,d} = 1$

1st Eigenvector: u_1

How good are empirical versions, $\hat{\lambda}_{1,d}, \cdots, \hat{\lambda}_{d,d}, \hat{u}_1$

as estimates?

UNC, Stat & OR

HDLSS Math. Stat. of PCA

Consistency (big enough spike):

For $\alpha > 1$,

$$Angle(u_1, \hat{u}_1) \to 0$$

Strong Inconsistency (spike not big enough):

For $\alpha < 1$,

$$Angle(u_1, \hat{u}_1) \to 90^\circ$$

UNC, Stat & OR

HDLSS Math. Stat. of PCA

Intuition: Random Noise $\sim d^{1/2}$

For $\alpha > 1$ (recall on scale of variance),

Spike pops out of *noise sphere*

Key Point:
Must do the Math

For $\alpha < 1$,

Spike contained in *noise sphere*

HDLSS Math. Stat. of PCA

Consistency of Eigenvalues?

$$\hat{\lambda}_{1,d} \xrightarrow{L} \frac{\chi^2_n}{n} \lambda_{1,d}$$

- Eigenvalues *Inconsistent*
- But Known Distribution
- Unless $n \to \infty$ as well

UNC, Stat & OR

Statistical Context of HDLSS Math

Careful look at:

- PCA Consistency - $\alpha > 1$ spike

Independent of Sample Size,

So true for $n = 1$ (!?!)

UNC, Stat & OR

Statistical Context of HDLSS Math

Careful look at:

- PCA Consistency - $\alpha > 1$ spike

Independent of Sample Size,

So true for $n = 1$ (!?!)

Absurd, shows assumption too strong

for practice ???

UNC, Stat & OR

Motivation of HDLSS Asy's

HDLSS

PCA

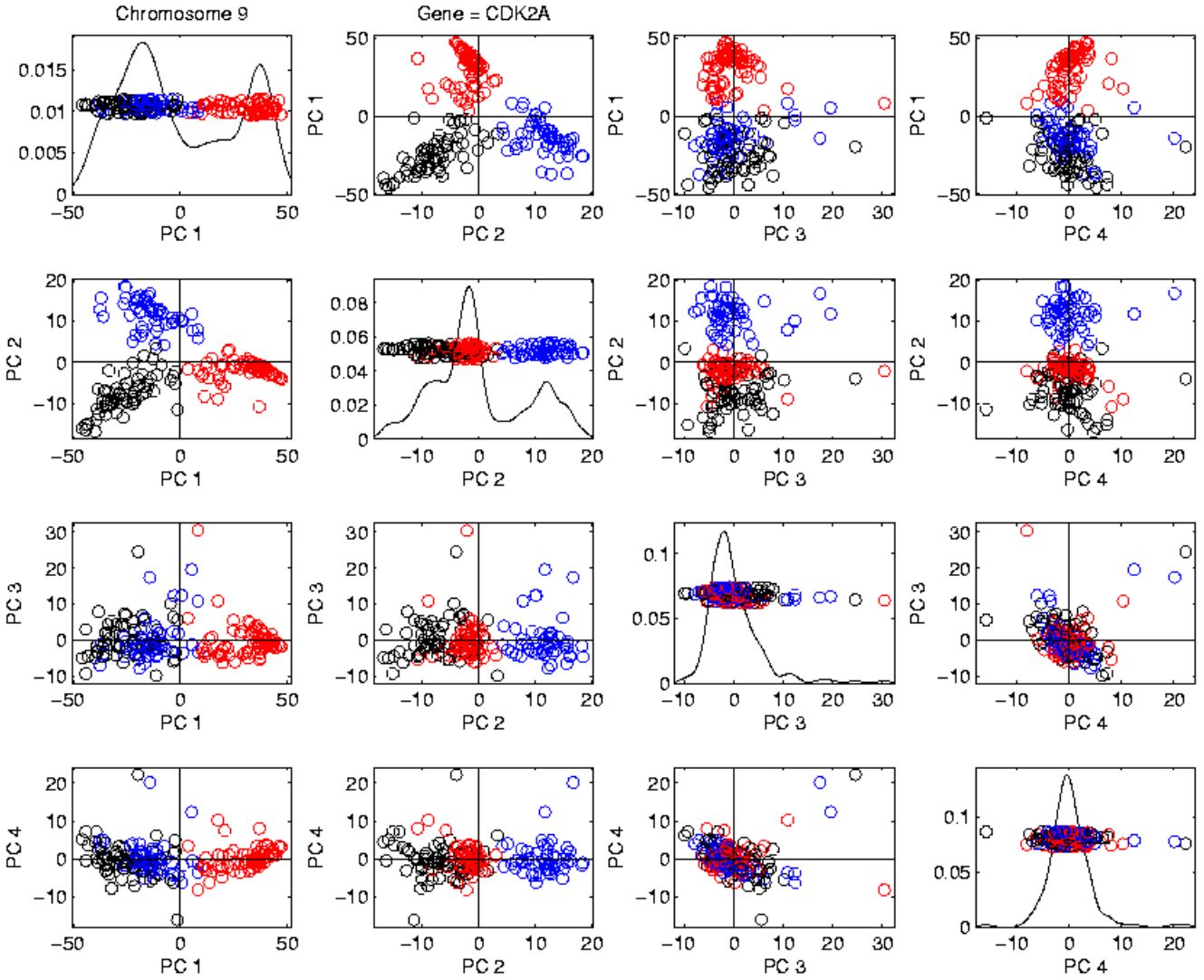
Often

Finds

Signal

Not *Pure*

Noise



UNC, Stat & OR

Statistical Context of HDLSS Math

Recall Theoretical Separation:

- Strong Inconsistency - $\alpha < 1$ spike
- Consistency - $\alpha > 1$ spike

Mathematically Driven Conclusion:

Real Data Signals Are This Strong

UNC, Stat & OR

Statistical Context of HDLSS Math

An Interesting Objection:

Should *not Study Angles* in PCA

Recall, for $\alpha > 1$,

$$Angle(u_1, \hat{u}_1) \to 0$$

For $\alpha < 1$,

$$Angle(u_1, \hat{u}_1) \to 90^\circ$$

Statistical Context of HDLSS Math

An Interesting Objection:

Should *not Study Angles* in PCA

Because PC Scores (i.e. projections)

Not Consistent

UNC, Stat & OR

Statistical Context of HDLSS Math

An Interesting Objection:

Should *not Study Angles* in PCA

Because PC Scores (i.e. projections)

Not Consistent

For Scores $\hat{s}_{i,j} = P_{\hat{v}_j} x_i$

What we study in PCA scatterplots

UNC, Stat & OR

Statistical Context of HDLSS Math

An Interesting Objection:

Should *not Study Angles* in PCA

Because PC Scores (i.e. projections)

Not Consistent

For Scores $\hat{s}_{i,j} = P_{\hat{v}_j} x_i$ and $s_{i,j} = P_{v_j} x_i$

Can Show $\frac{\hat{s}_{i,j}}{s_{i,j}} \to R_j \neq 1$ (Random!)

Thanks to Dan Shen

UNC, Stat & OR

Statistical Context of HDLSS Math

PC Scores (i.e. projections)

Not Consistent

So how can PCA find *Useful Signals* in Data?

UNC, Stat & OR

Motivation of HDLSS Asy's

HDLSS

PCA

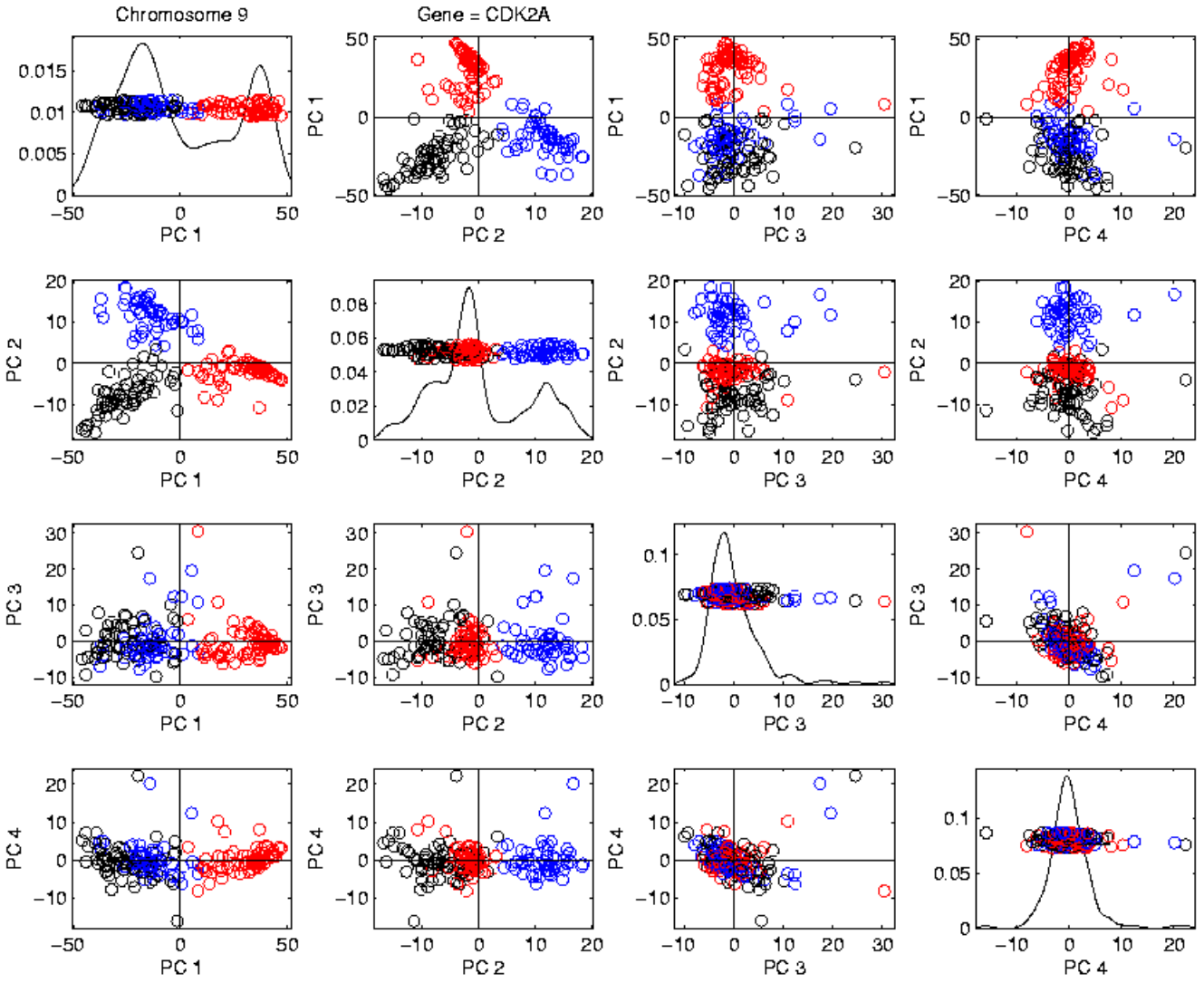
Often

Finds

Signal

Not *Pure*

Noise



UNC, Stat & OR

Statistical Context of HDLSS Math

PC Scores (i.e. projections)

Not Consistent

So how can PCA find *Useful Signals* in Data?

Key is "Proportional Errors" $\frac{\hat{s}_{i,j}}{s_{i,j}} \to R_j \neq 1$

Axes have *Inconsistent Scales*,

But Relationships are Still Useful

Key Point: Careful About Interpreting the Math

UNC, Stat & OR

Statistical Context of HDLSS Math

Direct Solution:

- Aoshima & Yata
- Consistent Scores Estimates
- Based on Dual Space (Gram Matrix) Ideas

UNC, Stat & OR

HDLSS Deep Open Problem

In PCA Consistency:

- Strong Inconsistency - $\alpha < 1$ spike
- Consistency - $\alpha > 1$ spike

What happens at boundary ($\alpha = 1$)???

UNC, Stat & OR

HDLSS Deep ~~Open Problem~~ Result

In PCA Consistency:

- Strong Inconsistency - $\alpha < 1$ spike
- Consistency - $\alpha > 1$ spike

What happens at boundary ($\alpha = 1$)???

∃ interesting Limit Distn's (Sen et al)

UNC, Stat & OR

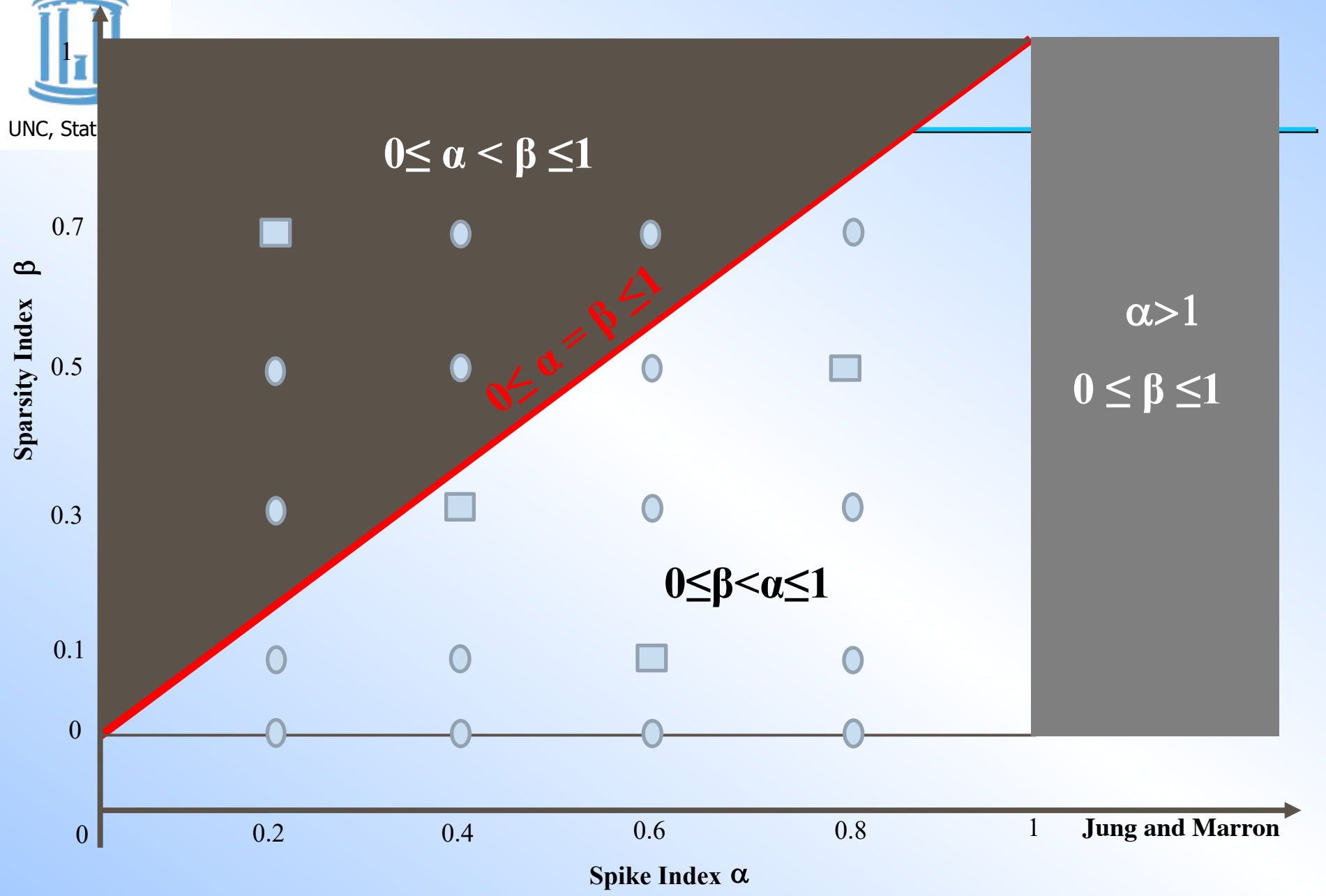
HDLSS & Sparsity

New Area Opened in PhD Work by Dan Shen

- Uses Spike Index, α
- And Sparsity Index, β
- Explores Consistency

& Strong Inconsistency

Sparsity Gives Broad New Region


1
0.7
0.5
0.3
0.1
0
Sparsity Index β
0
0.2
0.4
0.6
0.8
1
Jung and Marron
Spike Index α
0≤ α < β ≤1
0≤ α = β ≤1
α>1
0 ≤ β ≤1
0≤β<α≤1

UNC, Stat & OR

HDLSS & Other Asymptotics

Larger Context: PhD Work by Dan Shen

Explores PCA Consistency under all of:

Classical: d fixed, $n \to \infty$

Portnoy: $d, n \to \infty$, $d \ll n$

Random Matrices: $d, n \to \infty$, $d \sim n$

HDMSS: $d, n \to \infty$, $d \gg n$

HDLSS: $d \to \infty$, n fixed

UNC, Stat & OR

HDLSS & Other Asymptotics

Larger Context: PhD Work by Dan Shen

Explores PCA Consistency under all of:

Classical: d fixed, $n \to \infty$

Interesting and Surprising Case

Portnoy: $d, n \to \infty$, $d \ll n$

Random Matrices: $d, n \to \infty$, $d \sim n$

HDMSS: $d, n \to \infty$, $d \gg n$

HDLSS: $d \to \infty$, n fixed

UNC, Stat & OR

HDMSS

Asymptotics: $d, n \to \infty$ $d \gg n$

Leading Groups:

- Fan, et al (Princeton)
- Aoshima, Yata (Tsukuba)

UNC, Stat & OR

HDMSS, Fan View

Asymptotics: $d, n \to \infty$ $d \gg n$

"Ultra High Dimension" (Fan & Lv 2008):

1. Driver: $n \to \infty$

   (Classical Viewpoint)

2. Follower: $d \sim e^n$

   (Perhaps Impressive?)

UNC, Stat & OR

HDMSS, Aoshima View

Asymptotics: $d, n \to \infty$ $d \gg n$

1. Driver: $d \to \infty$

   (New Viewpoint)

2. Follower: $n \sim \log(d)$

   (Mathematically Equivalent?)

HDMSS, Personal Choice

Aoshima View:

1. Driver: $d \to \infty$
2. Follower: $n \sim \log(d)$

Since this allows easy interface with HDLSS:

$d \to \infty$, with n fixed

UNC, Stat & OR

HDLSS Asymptotics

Publishability???

PCA Consistency for $n = 1$
Inconsistent, but Useful Scores

- Basic Ideas Developed as "Concentration of Measure" by Talagrand (1990s)
- Many are Corollaries of "Concentration Inequalities", Massart, Lugosi, van der Vaart, Kolchinski, Tsybakov, …
- Value (∴ pub'able) of HDLSS asymptotics:

Statistical Insights

UNC, Stat & OR

HDLSS Asymptotics

Personal Observations:

HDLSS world is…

- Surprising (many times!)

  [Think I've got it, and then …]

- Mathematically Beautiful (?)
- Practically Relevant
- Publishable???

Key Point:
Must do the Math

UNC, Stat & OR

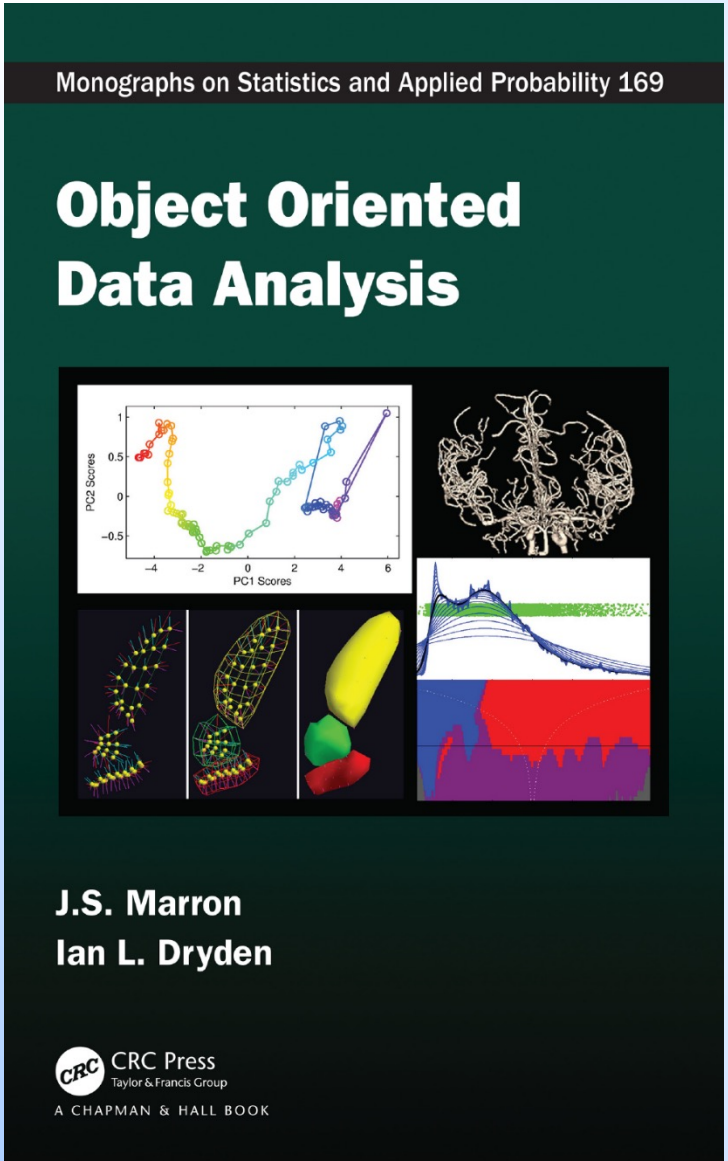
The Future of HDLSS Asymptotics?

1. Address your favorite statistical problem…
2. HDLSS versions of classical optimality results?
3. Continguity Approach (~Random Matrices)
4. Rates of convergence?
5. Improved Discrimination Methods?

It is early days…

UNC, Stat & OR

My Current Research



UNC, Stat & OR

My Current Research

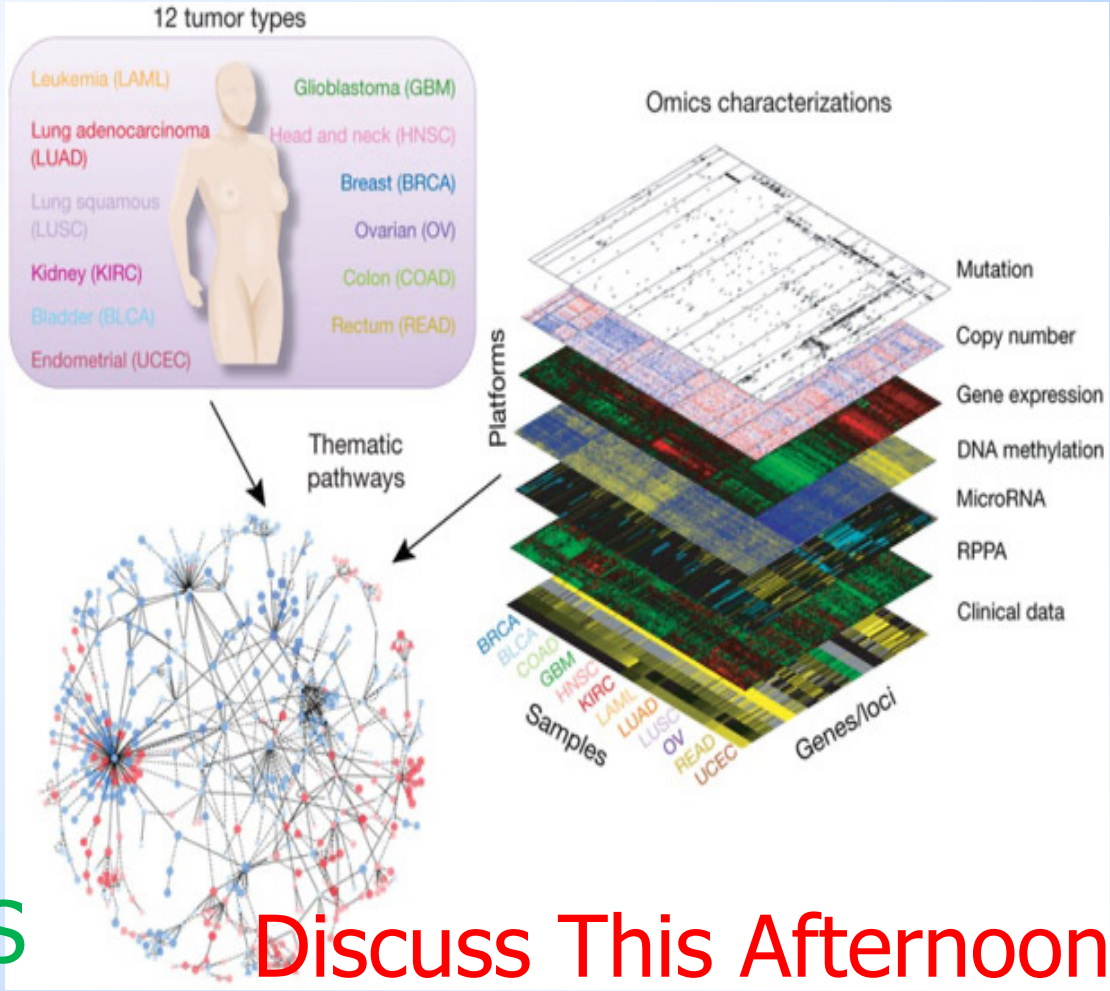
Data Integration:

Important

New Methods:

JIVE

DIVAS



Interesting HDLSS Asy. Challenges

Discuss This Afternoon

Figure : The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013)
The Cancer Genome Atlas Pan-Cancer analysis project.