Handling Heterogeneity in Big Data

J. S. Marron

University of North Carolina

November 16, 2015
Special Thanks

Department of Statistics and Applied Prob.
National University of Singapore
For Many Discussions \rightarrow This Talk
Comments on Current “Fashion”

Big Data

• Isn’t It Just Statistics?
• Yes, But We Need to Remind Folks
• Maybe Bigger Challenge:

Complex Data
Comments on Current “Fashion”

Complex Data

Important Viewpoint:

Object Oriented Data Analysis
Comments on Current “Fashion”

Big Data

• Many Folks Want to Play
  • Computer Science / E. E.
  • (Applied) Mathematics
  • Bioinformatics
  • Others (All Sciences are Collecting …)
Comments on Current “Fashion”

Big Data

• What Does Statistics Have to Offer?
Comments on Current “Fashion”

Big Data

• What Does Statistics Have to Offer?
  • Vast Collective Knowledge

(Efficiency: Avoid Massive Re-Invention)
Comments on Current “Fashion”

Big Data

• What Does Statistics Have to Offer?
  • Vast Collective Knowledge
  • Broad Experience with Pitfalls

(Efficiency: Avoid Well Known Traps)
Comments on Current “Fashion”

Big Data

• What Does Statistics Have to Offer?
  • Vast Collective Knowledge
  • Broad Experience with Pitfalls
  • Diverse, Well Developed Approaches
Comments on Current “Fashion”

Big Data

• What Does Statistics Have to Offer?
  • Vast Collective Knowledge
  • Broad Experience with Pitfalls
  • Diverse, Well Developed Approaches

{Need to Keep Reminding Everybody}
Comments on Current “Fashion”

Big Data

- How Should Statistics Get Engaged?
Comments on Current “Fashion”

Big Data

- How Should Statistics Get Engaged?
- Major Tasks / Goals:
  - Parallelization
  - More Appropriate Models / Concepts
Challenge from the Recent Media:

Mayer-Schönberger and Cukier (2014)

“Big Data: A Revolution That Will Transform How We Live, Work, and Think”
Challenge from the Recent Media:

Mayer-Schönberger and Cukier (2014)

Major Premise: Differing Data Analytic Goals
"Correlational" vs. "Causal"
“Causal” Data Analysis:

- Goal: Underlying Causes of Phenomena
- Approach: Classical “Scientific Method”
  - Formulate Hypothesis
  - Collect Data
  - Test Hypothesis
- Consequences:
  Solid Knowledge w/ Measurable Certainty
“Correlational” Data Analysis:

- **Goal:** Find (and Use) Mere Correlations
- **Motivation:** Correlations are
  - Useful (e.g. ___ Recognition Software)
  - Valuable (Buying and Selling of Data…)
  - Insightful?????
- **Consequences:**
  
  **Automatic** Solutions to Some Hard Problems
Correlation vs. Causation

How New Is This Discussion?
Correlation vs. Causation

How New Is This Discussion?

Naïve Readers [Of Mayer-Schönberger and Cukier (2014)]:

This is Exciting!!!
Great New Ideas!!!
Change Statistics Curricula!!!
Start Up “Data Analytics”!!!
Correlation vs. Causation

How New Is This Discussion?

Statistics
Correlation vs. Causation

How New Is This Discussion?

Pattern Recognition
Artificial Intelligence
Neural Networks
Data Mining
Machine Learning
Correlation vs. Causation

How New Is This Discussion?

Statistics

Time

Pattern Recognition
Artificial Intelligence
Neural Networks
Data Mining
Machine Learning
???
Correlation vs. Causation

How New Is This Discussion?

Statistics

Pattern Recognition
Artificial Intelligence
Neural Networks
Data Mining
Machine Learning
Big Data
Correlation vs. Causation

How New Is This Discussion?

Some Came With Major New Ideas

Pattern Recognition
Artificial Intelligence
Neural Networks
Data Mining
Machine Learning
Big Data
**Correlation vs. Causation**

How New Is This Discussion?

Less So For Others, But More Focus On

Pattern Recognition
Artificial Intelligence
Neural Networks
Data Mining
Machine Learning
Big Data
Correlation vs. Causation

How New Is This Discussion?

Data Mining

Great Correlational Discovery
Correlation vs. Causation

How New Is This Discussion?

Data Mining

Great Correlational Discovery:

Super Market Scanner Data

Baby Diapers (aka Nappies) & Beer
Correlation vs. Causation

How New Is This Discussion?

Data Mining

Baby Diapers (aka Nappies) & Beer

Some Perspective:

- Correlational Discovery
- Makes Causational Sense

(Too Soon To Totally Dump Causation)
Correlation vs. Causation

Relative Emphasis???
Relative Emphasis???

Classical Statistics:

Correlation vs. Causation
Correlation vs. Causation

Relative Emphasis???

Mayer-Schönberger and Cukier:

Correlation vs. Causation
Correlation vs. Causation

Relative Emphasis???

Suggested Actual Future Course:

Correlation & Causation
Correlation vs. Causation

Relative Emphasis???

Suggested Actual Future Course:

Correlation & Causation

Note: Changes Are Needed in Curricula, Etc.
Comments on Current “Fashion”

Big Data

- How Should Statistics Get Engaged?
- Major Tasks / Goals:
  - Parallelization
  - More Appropriate Models / Concepts
Big Data Conceptual Models

Standard Generic Statistical Model:

Data are $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$

Time Tested Reasons:

- Central Limit Theorem
Big Data Conceptual Models

Standard Generic Statistical Model:

Data are \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \)

Time Tested Reasons:

- Central Limit Theorem
- Sum of Many, Small, \( \sim \) Indep’nt RVs
- Useful Model in Many Situations
Usual Multivariate Version:

Data are \( X_1, \ldots, X_n \sim N_d (\mu, \Sigma) \)
Aside on Terminology

Personal Notation:

- Dimension: $d$
- Sample size $n$
Aside on Terminology

Personal Notation:

- Dimension: \( d \)
- Sample size \( n \)

Why not \( p \)?
Aside on Terminology

Personal Notation:

- Dimension:  $d$
- Sample size  $n$

- Why not $p$? What does $p$ stand for?
  
  (Parameters??? Predictors???)
Aside on Terminology

Personal Notation:

- Dimension: $d$
- Sample size $n$

- Why not $p$? What does $p$ stand for? (Parameters??? Predictors???)

- Only Because of Statistical Tradition…

- Seems Strange Outside of Statistics
Usual Multivariate Version:

Data are $X_1, \ldots, X_n \sim N_d (\mu, \Sigma)$

Caution: Makes Critical Assumption

*Common* Data Generation Mechanism

“Homogeneous Data”
Common Data Generation Mechanism
“Homogeneous Data”

Terminology: Bühlmann & Meinshausen
(2014 ArXiv)
Big Data Conceptual Models

Common Data Generation Mechanism

“Homogeneous Data”

Terminology: Bühlmann & Meinshausen

Fundamental Observation:
Not Realistic for Many Big Data Contexts
Big Data Conceptual Models

*Common* Data Generation Mechanism

“Homogeneous Data”

**Not** Realistic for Many Big Data Contexts

Because of Their Origins:

- Concatenations of Smaller Data Sets
- Often From Multiple Experiments
- Across Investigators / Laboratories
Big Data Conceptual Models

Suggested New Viewpoint

“Heterogeneous Data”
Suggested New Viewpoint

“Heterogeneous Data”

Better Thought Model: Gaussian Mixtures

\[ X_1, \ldots, X_n \sim \sum_{j=1}^{k} w_j N_d \left( \mu_j, \Sigma_j \right) \]
Big Data Conceptual Models

Thought Model: Gaussian Mixtures

\[ X_1, \ldots, X_n \sim \sum_{j=1}^{k} w_j N_d (\mu_j, \Sigma_j) \]

Can Effectively Represent:

- Data Combined Across Laboratories
- Operator / Timing Effects
Thought Model: Gaussian Mixtures

\[ X_1, \ldots, X_n \sim \sum_{j=1}^{k} w_j N_d \left( \mu_j, \Sigma_j \right) \]

Can Effectively Represent:

- Data Combined Across Laboratories
- Operator / Timing Effects

Can be Handled With Careful Design
Thought Model: Gaussian Mixtures

\[ X_1, \ldots, X_n \sim \sum_{j=1}^{k} w_j N_d(\mu_j, \Sigma_j) \]

Can Effectively Represent:
- Data Combined Across Laboratories
- Operator / Timing Effects
- Biological Diversity (e.g. Male-Female or others...)
Thought Model: Gaussian Mixture Models

\[ X_1, \ldots, X_n \sim \sum_{j=1}^{k} w_j N_d \left( \mu_j, \Sigma_j \right) \]

Standard Approach:

Parameter Estimation

(∃ a large literature)
Big Data Conceptual Models

Thought Model: Gaussian Mixtures

\[ X_1, \ldots, X_n \sim \sum_{j=1}^{k} w_j N_d \left( \mu_j, \Sigma_j \right) \]

Standard Approach: Parameter Estimation

Big Data Challenge: Too Many Parameters to Effectively Estimate
Thought Model:
Gaussian Mix’s but *No* Parameter Estimation
Big Data Conceptual Models

Thought Model:
Gaussian Mix’s but No Parameter Estimation

Approaches:

- Identify with *Data Visualization*, …
Thought Model:
Gaussian Mix’s but no Parameter Estimation

Approaches:

- Identify with *Data Visualization*, …
- Methods Robust Against Unknown Mix’s
Robust Statistics

Potential New Area:

Robustness Against Unknow-able Mixtures
Meinshausen & Bühlmann (2014)

In Regression Settings:

“Maximin Aggregation”
Visualization of Heterogeneity

For High Dimensions,

Useful Approach:

Principal Component Analysis
Visualization of Heterogeneity

For High Dimensions,

Useful Approach:

Principal Component Analysis

Visualization is *More Important* than

“Dimension Reduction”
Visualization of Heterogeneity

For High Dimensions,

Useful Approach: Principal Component Analysis

Visualization is *More Important* than “Dimension Reduction”

Choose $d^*$, so that $\sum_{j=1}^{d^*} \lambda_j \approx 9\%$
Visualization of Heterogeneity

Informative PCA Data View

- Scatterplot Matrix
  - a.k.a. Draftsman’s Plot

Think “Front – Side – Top View”  3d → 2d
Visualization of Heterogeneity

Interesting Example:

Breast Cancer Microarray Data

Visualization of Heterogeneity

Interesting Example:

Breast Cancer Microarray Data


Records “Activity” of Each Gene

\[ n = 107, \quad d = 5961 \]
Heterogeneity in Microarray Data

PCA Scatterplot View

Stanford Breast Cancer Data

Raw Data - Source Colored

Raw PC Axes
Heterogeneity in Microarray Data

PCA
Scatterplot View
1-d Projections On Diagonal
(jitter plot & KDE)
Heterogeneity in Microarray Data

PCA
Scatterplot View

2-d Projections
Off Diagonal
Heterogeneity in Microarray Data

PCA
Scatterplot View

Note Common Axes
Heterogeneity in Microarray Data

PCA Scatterplot View

Note Common Axes
Heterogeneity in Microarray Data

PCA
Scatterplot
View
Think:
Front
Side
Top
Rotations
Heterogeneity in Microarray Data

Show Known Heterogeneity Using Colors
Heterogeneity in Microarray Data

RNA Source Colors
(Ink Used In Chip Prints)
Heterogeneity in Microarray Data

RNA Source Colors

Trouble in PC1, Dir’n of Maximal Variation!!!
Heterogeneity in Microarray Data

Print Batch Colors (Time of Printing)
Heterogeneity in Microarray Data

Print Batch Colors

Shows in PC4
Heterogeneity in Microarray Data

Orthogonality Of Effects???

PC1 vs. PC4

Well Designed (Randomized) Experiment!
Heterogeneity in Microarray Data

Serious Question: Anything Useful in Data?
Heterogeneity in Microarray Data

Color by Cancer Classes

PC2 & PC3 Looks Good
Heterogeneity in Microarray Data

Color by Cancer Classes

Major Discovery: Luminal A vs. Luminal B
Batch Effect Adjustment

Goal: Remove Heterogeneity
Batch Effect Adjustment

Recall

Ink (RNA) Effect
Batch Effect Adjustment

Recall

Time (Made)

Effect
Batch Effect Adjustment

Recall

Class Colors
Batch Effect Adjustment

Turn Class Colors Into Symbols
Batch Effect Adjustment

So Still See Biological Classes
Batch Effect Adjustment

Recall RNA (I nk) Effect
Batch Effect Adjustment

Replace PC 4 with:

DWD Direction,

Gives “Best” Separation Of Groups
Batch Effect Adjustment

Slide Groups Together (in Full Space) Along DWD Direction

Classes Merge Together Well
Batch Effect Adjustment

Stanford Breast Cancer Data

Source Adj. = Source Colored

Raw PC Axes

+ LUMA, x LUMB, o Norm, v ERBB2, Basal
Batch Effect Adjustment

Check Good Merging of Classes
Batch Effect Adjustment

Recall Time (Fabrication Batch) Effect

- Stanford Breast Cancer Data
- Source Adj. – Batch Colored
- Raw PC Axes
- LumA, x LumB, o Norm, v ERBB2, Basal
Batch Effect Adjustment

Add PC 5, Since Don’t Want to Replace PC 4 With DWD Direction
Batch Effect Adjustment

Blue vs. Green ∪ Red

DWD Direction
Batch Effect Adjustment

Eliminate That Component
Batch Effect Adjustment

Red vs.
Green \cup Blue
DWD Direction
Batch Effect Adjustment

Eliminate That Component
Batch Effect Adjustment

Put PC 5 Back
Batch Effect Adjustment

Return to First 4 PCs
Batch Effect Adjustment

Show Class Colors of Adjusted Data
Batch Effect Adjustment

Recompute PCA (i.e. Rotate Axes)

Now Classes (Biological) Dominate First 3 PCs
DWD Background

Recall:
Direction
of Good
Separation
Recall: Direction of Good Separation

How Does It Work?
DWD Background

2-d Toy Data

“Best” Separating Direction???
DWD Background

Mean Difference (Centroid)
DWD Background

Distance Weighted Discrimination

\[
\min_{\text{dir}} \sum_{i=1}^{n} \frac{1}{r_i}
\]
Distance
Weighted
Discrimination

\[
\min_{\text{dir}} \sum_{i=1}^{n} \frac{1}{r_i}
\]

Separating Plane

Gaussian 2-d Toy Data: DWD Direction
DWD Background

Distance
Weighted
Discrimination

\[ \min_{\text{dir}} \sum_{i=1}^{n} \frac{1}{r_i} \]

DWD Direction is Normal Vector
DWD Robustness to Heterogeneity

Why Not Use Means for Heterogeneity Adjustment?

(DWD Is Much More Complicated)
Why Not Use Means for Heterogeneity Adjustment?

(DWD is Much More Complicated)
DWD Robustness to Heterogeneity

Why Not Use Means for Heterogeneity Adjustment?

(DWD is Much More Complicated)

Bioinformatician’s (early) View:

DWD Works Like Magic
(On Real Data)
DWD Robustness to Heterogeneity

Why Not Use Means for Heterogeneity Adjustment?

(DWD is Much More Complicated)

Key Concept (Xuxin Liu):

Unbalanced, Unknown Subtypes
DWD Robustness to Heterogeneity

Why Not Use Means for Heterogeneity Adjustment?

(DWD is Much More Complicated)

Key Concept (Xuxin Liu):

Unbalanced, Unknown Subtypes

Robustness Against Unknown Heterogeneity
DWD Robustness to Heterogeneity

2-d Toy Example

Balanced Mixture
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture

(Diminishing Discriminatory Power)
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture
DWD Robustness to Heterogeneity

Note: Are Adjusting For Known Heterogeneity
DWD Robustness to Heterogeneity

Note: Are Adjusting For Known Heterogeneity In Presence Of Unknown Heterogeneity
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture

Note: Losing Distinction To Be Studied
DWD Robustness to Heterogeneity

2-d Toy Example

Unbalanced Mixture
DWD Robustness to Heterogeneity

Why Not Use Means for Heterogeneity Adjustment?

(DWD is Much More Complicated)

Key Concept (Xuxin Liu):

Unbalanced, Unknown Subtypes

Robustness Against

Unknown Heterogeneity
Overview:

MD: Not Robust: Inaccurate Adjustment

SVM: Robust, But Poor Adjustment & Vis’n

DWD: Robust, Good Adjustment & Vis’n (Balance Between Extremes)
DWD Visualization

Another Important Application of DWD
DWD Visualization

Important Facts about PCA:

- Strong Track Record of Great Visualization
- But Only Feels Variation
- Not Class Labels
- Thus Can Miss Important Data Aspects
DWD Visualization

NCI 60 Data

Interesting Microarray Example

Diverse Cancer Types
DWD Visualization

NCI 60 Data
NSC Lung
Colon
Breast
Ovarian
Leukemia
Renal
Melanoma
DWD Visualization

NCI 60 Data

See Types?
NCI 60 Data

See Types?

Melanoma
NCI 60 Data

See Types?

Melanoma

Leukemia
DWD Visualization

NCI 60 Data

See Types?

Melanoma

Leukemia

Others???
DWD Visualization

NCI 60 Data
Others Types
Less Clear

But Only a 4-d View

Need More?
NCI 60 Data

Extended PC Components
NCI 60 Data
Extended PC Components:
Still PC1 – 4
DWD Visualization

NCI 60 Data

Extended PC Components: Still PC1 – 4

But vs. PC 5 – 8
NCI 60 Data
Now See
Renal Type
NCI 60 Data
Now See
Renal Type
But No Other
Types
(Also in Higher
dim’s)
Recall Important PCA Property:

Feels Variation, NOT Class Labels

Improved View: DWD
NCI 60 Data
Using DWD Directions
Note: Classes Much More Distinct
DWD Visualization

NCI 60 Data

DWD Trained on Renal vs. CNS

(but project all others)
DWD Visualization

NCI 60 Data

DWD Trained on Ovar vs. NSCLC
DWD Visualization

NCI 60 Data

DWD Trained on
Leuk vs. Colon
DWD Visualization

NCI 60 Data

DWD Trained on
Melan vs. Breast
Main Point:

When Class Labels are Known, DWD Is Effective At Highlighting Differences
Many Open Theoretical Problems

DWD Applications:
- Batch Adjustment
- Visualization
- Hypothesis Testing

Valuable Property:
Robustness Against Unknown Heterog’ty

Major New Math. Stat. Challenge:
Precise Formulation & Better Methods
Some Take-Away Concepts

- Big Data: Correlation vs. Causation
- Major Challenge: Data Heterogeneity
- Goal: Robustness Against Unknown Mixtures
- Regression: Maximin Aggregation
- Batch Adjustment: DWD
- DWD Also Hetero-Robust For:
  - Classification
  - Visualization
  - Hypothesis Testing