Statistical Data Mining
A Global View and Some Research Potentials

Dennis Lin
DKL5@psu.edu
The Pennsylvania State University

01 November, 2004
Institute for Computational Science
What Am I?

- Bachelor Degree in (pure) Mathematics
- PhD in Statistics
- Professor in Supply Chain Management (Business School)
- Professor in Statistics, Adjunct (College of Science)
- Professor in Industrial Engineering, Adjunct (College of Engineering)
- Advisor, IRS (USA) & Bureau of Statistics (Government)
Data Collection/Data Quality
- Sampling Survey
- Design of Experiment
- Population Data

Data Analysis
- $n < p$ or $n << p$
- Very large $n$
- Very large $p$
- HDLSS: High Dimension Low Sample Size (Marron)

You can analyze data as many different ways as you want, but you collect data only once!
What is your favorite symbol?

- Chemist
- Mathematician
- Bayesian
- Accountant
- Physicist
- Statistician
- B-school
- Secretary

Symbols:
- $\equiv$
- $\leq$
- $\infty$
- $\otimes \oplus$
- $\approx$
- $\sim$
- $\$
Olympic—Who’s fastest in the world?

- 100 meters
  - Hurdle race
  - Marathon race
  - Walking race
- Swimming
  - 100 meters
  - figure skating
  - diving
- Weight-lifting
Data Mining is more useful in empirical study & experience accumulation, namely “induction” type.
Brief Historical Development

Optimization (SA, GA, Neural Networks etc)
Statistics (classification tree, projection pursuit)

Data Mining (data-base)
Artificial Intelligent (rule-base)

Knowledge Discovery

Machine Learning

Statistical Learning/Math Learning etc.
Data Mining

- **Data**
  (re-design and maintain existing database)

- **Mining**
  (Analysis) -- our focus

Statistical Data Mining
Statistics

Making Sense Out of Numbers
What is Data Mining?

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data.

Viewed as part of the Knowledge Discovery process.

Uses tools from Computer Science and Artificial Intelligence as well as Statistics.
Pattern Findings

Best: New & Useful findings
Better: New & Not so useful findings
Good: Nothing
Worse: False Negative findings
Worst: False Positive findings
There are so many new and correct in your paper.

Unfortunately, what is new is not correct and what is correct is not new.

My brain has two parts: one is left, and another is right.

The left brain has nothing right; and the right brain has nothing left.
Titanic Data
(a total of 2208 cases)

<table>
<thead>
<tr>
<th>Survived</th>
<th>Age</th>
<th>Gender</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>C</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>M</td>
<td>3</td>
</tr>
</tbody>
</table>
# The Titanic Data

## Age By Survived

<table>
<thead>
<tr>
<th>Count Col % Row %</th>
<th>D</th>
<th>S</th>
<th></th>
<th>Count Col % Row %</th>
<th>D</th>
<th>S</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1438</td>
<td>661</td>
<td>2099</td>
<td>F</td>
<td>126</td>
<td>344</td>
<td>470</td>
</tr>
<tr>
<td></td>
<td>96.51</td>
<td>92.06</td>
<td>31.49</td>
<td>8.46</td>
<td>47.91</td>
<td>73.19</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>52</td>
<td>57</td>
<td>109</td>
<td>M</td>
<td>1364</td>
<td>374</td>
<td>1738</td>
</tr>
<tr>
<td></td>
<td>3.49</td>
<td>7.94</td>
<td>52.29</td>
<td>91.54</td>
<td>52.09</td>
<td>21.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1490</td>
<td>718</td>
<td>2208</td>
<td>1490</td>
<td>718</td>
<td>2208</td>
<td></td>
</tr>
</tbody>
</table>

## Gender By Survived

<table>
<thead>
<tr>
<th>Count Col % Row %</th>
<th>D</th>
<th>S</th>
<th></th>
<th>Count Col % Row %</th>
<th>D</th>
<th>S</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>126</td>
<td>344</td>
<td>470</td>
<td>122</td>
<td>203</td>
<td>673</td>
<td>219</td>
</tr>
<tr>
<td></td>
<td>8.46</td>
<td>47.91</td>
<td>73.19</td>
<td>8.19</td>
<td>28.27</td>
<td>45.17</td>
<td>30.50</td>
</tr>
<tr>
<td></td>
<td>167</td>
<td>118</td>
<td>285</td>
<td>528</td>
<td>178</td>
<td>673</td>
<td>219</td>
</tr>
<tr>
<td></td>
<td>11.21</td>
<td>16.43</td>
<td>41.40</td>
<td>35.44</td>
<td>24.79</td>
<td>74.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1490</td>
<td>718</td>
<td>2208</td>
<td>1490</td>
<td>718</td>
<td>2208</td>
<td></td>
</tr>
</tbody>
</table>

## Class By Survived

<table>
<thead>
<tr>
<th>Count Col % Row %</th>
<th>D</th>
<th>S</th>
<th></th>
<th>Count Col % Row %</th>
<th>D</th>
<th>S</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>122</td>
<td>203</td>
<td>325</td>
<td>673</td>
<td>219</td>
<td>892</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.19</td>
<td>28.27</td>
<td></td>
<td>45.17</td>
<td>30.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>167</td>
<td>118</td>
<td>285</td>
<td>74.79</td>
<td>24.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.21</td>
<td>16.43</td>
<td></td>
<td>75.45</td>
<td>25.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>528</td>
<td>178</td>
<td>706</td>
<td>673</td>
<td>178</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35.44</td>
<td>24.79</td>
<td></td>
<td>45.17</td>
<td>25.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crew</td>
<td>1490</td>
<td>718</td>
<td>2208</td>
<td>1490</td>
<td>718</td>
<td>2208</td>
<td></td>
</tr>
</tbody>
</table>
### Gender By Survived

<table>
<thead>
<tr>
<th>Count</th>
<th>Col %</th>
<th>Row %</th>
<th>D</th>
<th>S</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F</strong></td>
<td>126</td>
<td>8.46</td>
<td>26.81</td>
<td>344</td>
<td>47.91</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>1364</td>
<td>91.54</td>
<td>78.48</td>
<td>374</td>
<td>52.09</td>
</tr>
</tbody>
</table>

**Total:** 1490 718 2208

- $\chi^2$ Test
- **Odd Ratio Test**

DeVeaux
Tree model

M

F

Adult

Child

2 or 3

1 or Crew

1 or 2

3

3

14%

Crew

23%

1st

27%

1st

93%

1,2,C

46%
Complexity

- Data Complexity
- Algorithm Complexity

Wegman
### The Huber-Wegman Taxonomy of Data Set Sizes

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Data Set Size in Bytes</th>
<th>Storage Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiny</td>
<td>$10^2$</td>
<td>Piece of Paper</td>
</tr>
<tr>
<td>Small</td>
<td>$10^4$</td>
<td>A Few Pieces of Paper</td>
</tr>
<tr>
<td>Medium</td>
<td>$10^6$</td>
<td>A Floppy Disk</td>
</tr>
<tr>
<td>Large</td>
<td>$10^8$</td>
<td>Hard Disk</td>
</tr>
<tr>
<td>Huge</td>
<td>$10^{10}$</td>
<td>Multiple Hard Disks</td>
</tr>
<tr>
<td>Massive</td>
<td>$10^{12}$</td>
<td>Robotic Magnetic Tape Storage Silos</td>
</tr>
<tr>
<td>Supermassive</td>
<td>$10^{15}$</td>
<td>Distributed Data Archives</td>
</tr>
</tbody>
</table>
Algorithmic Complexity

$O(n^{1/2})$  Plot a Scatterplot

$O(n)$  Calculate Means, Variances, Kernel Density Estimates

$O(n \log(n))$  Calculate Fast Fourier Transforms

$O(nc)$  Calculate Singular Value Decomposition of an $r \times c$ Matrix; Solve a Multiple Linear Regression

$O(n^2)$  Solve most Clustering Algorithms

$O(a^n)$  Detect Multivariate Outliers
**Complexity**

*Number of Operations for Algorithms of Various Computational Complexities and Various Data Set Sizes*

<table>
<thead>
<tr>
<th></th>
<th>$n^{1/2}$</th>
<th>$n$</th>
<th>$n \log(n)$</th>
<th>$n^{3/2}$</th>
<th>$n^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiny</td>
<td>10</td>
<td>$10^2$</td>
<td>2x$10^2$</td>
<td>10$^3$</td>
<td>10$^4$</td>
</tr>
<tr>
<td>small</td>
<td>10$^2$</td>
<td>$10^4$</td>
<td>4x$10^4$</td>
<td>10$^6$</td>
<td>10$^8$</td>
</tr>
<tr>
<td>medium</td>
<td>10$^3$</td>
<td>$10^6$</td>
<td>6x$10^6$</td>
<td>10$^9$</td>
<td>10$^{12}$</td>
</tr>
<tr>
<td>large</td>
<td>10$^4$</td>
<td>$10^8$</td>
<td>8x$10^8$</td>
<td>10$^{12}$</td>
<td>10$^{16}$</td>
</tr>
<tr>
<td>huge</td>
<td>10$^5$</td>
<td>$10^{10}$</td>
<td>$10^{11}$</td>
<td>10$^{15}$</td>
<td>10$^{20}$</td>
</tr>
</tbody>
</table>

Wegman
**Complexity**

Computational Feasibility on a Pentium PC  
10 megaflop performance assumed  

<table>
<thead>
<tr>
<th>$n$</th>
<th>$n^{1/2}$</th>
<th>$n$</th>
<th>$n \log(n)$</th>
<th>$n^{3/2}$</th>
<th>$n^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiny</td>
<td>$10^6$ seconds</td>
<td>$10^5$ seconds</td>
<td>$2 \times 10^{-5}$ seconds</td>
<td>$.0001$ seconds</td>
<td>$.001$ seconds</td>
</tr>
<tr>
<td>small</td>
<td>$10^5$ seconds</td>
<td>$.001$ seconds</td>
<td>$.004$ seconds</td>
<td>$.1$ seconds</td>
<td>10 seconds</td>
</tr>
<tr>
<td>medium</td>
<td>$.0001$ seconds</td>
<td>$.1$ seconds</td>
<td>$.6$ seconds</td>
<td>1.67 minutes</td>
<td>1.16 days</td>
</tr>
<tr>
<td>large</td>
<td>$.001$ seconds</td>
<td>10 seconds</td>
<td>1.3 minutes</td>
<td>1.16 days</td>
<td>31.7 years</td>
</tr>
<tr>
<td>huge</td>
<td>$.01$ seconds</td>
<td>16.7 minutes</td>
<td>2.78 hours</td>
<td>3.17 years</td>
<td>317,000 years</td>
</tr>
</tbody>
</table>

Wegman
### Complexity

#### Computational Feasibility on a Silicon Graphics Onyx Workstation

300 megaflop performance assumed

<table>
<thead>
<tr>
<th>Size</th>
<th>( n )</th>
<th>( n^{1/2} )</th>
<th>( n )</th>
<th>( n \log(n) )</th>
<th>( n^{3/2} )</th>
<th>( n^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiny</td>
<td>( 3.3 \times 10^{-8} ) seconds</td>
<td>( 3.3 \times 10^{-7} ) seconds</td>
<td>( 6.7 \times 10^{-7} ) seconds</td>
<td>( 3.3 \times 10^{-6} ) seconds</td>
<td>( 3.3 \times 10^{-5} ) seconds</td>
<td></td>
</tr>
<tr>
<td>small</td>
<td>( 3.3 \times 10^{-7} ) seconds</td>
<td>( 3.3 \times 10^{-5} ) seconds</td>
<td>( 1.3 \times 10^{-4} ) seconds</td>
<td>( 3.3 \times 10^{-3} ) seconds</td>
<td>( .33 ) seconds</td>
<td></td>
</tr>
<tr>
<td>medium</td>
<td>( 3.3 \times 10^{-6} ) seconds</td>
<td>( 3.3 \times 10^{-3} ) seconds</td>
<td>( .02 ) seconds</td>
<td>( 3.3 ) seconds</td>
<td>( 55 ) minutes</td>
<td></td>
</tr>
<tr>
<td>large</td>
<td>( 3.3 \times 10^{-5} ) seconds</td>
<td>( .33 ) seconds</td>
<td>( 2.7 ) seconds</td>
<td>( 55 ) minutes</td>
<td>( 1.04 ) years</td>
<td></td>
</tr>
<tr>
<td>huge</td>
<td>( 3.3 \times 10^{-4} ) seconds</td>
<td>( 33 ) seconds</td>
<td>( 5.5 ) minutes</td>
<td>( 38.2 ) days</td>
<td>( 10,464 ) years</td>
<td></td>
</tr>
</tbody>
</table>

Wegman
Statistical Data Mining

Need Statistical methodologies/algorithms that is *computable* (under the constraints of computer memory and complexity).

So

- All Statistical methodologies need to be labeled its complexity.
- For powerful $O(n^2)$ methodologies, an approximate $O(n)$ algorithm is needed.
Problems

- Classification (Supervised Learning)
- Clustering (Unsupervised Learning)
- Pattern Recognition
- Association (Correlation)
- Modeling
- Estimation
- Prediction
- Description
- Visualization
- Etc.
Some Examples

- Barclaycard (UK Credit Card Company)
  - 350 million transactions a year
- Walmart (USA Retailer)
  - 7 billion transactions a year
- AT&T
  - 70 billion long distant calls per year
- Mobil Oil
  - 100 Terabytes of data (oil exploration)
- NASA
  - 50 Gigabytes data per hour
- Human Genome Project
More Examples

- AT&T Telephone Fraud Usage Detection
- AT&T Internet Traffic Data
- Lucent TSC “Trouble Ticket” Database
- Grocery Store
- Credit Card Company
- Risk of Diseases
- E-commerce
- Drug Discovery
Case Study—British Saveway

- Revenue—more than 10 billion$
- 34 type services
- 3rd largest chain stores in British
- 4TB data (weekly)
British Saveway

- Data Collection
  - Developing Credit Cards (500GB weekly)

- Automatic Statistical Analysis
  - Association
  - Classification, etc

- Modeling/Pattern Recognition
  - Specific product ranked 209 was purchased by highly valued customers

- Other Observations
  - Eight OJ (out of 28) are particularly popular
  - Sequence Discovery (prediction, mailing etc)
Recent Projects Involved

- Chinese Medicine Study (Fang)
- Teaching Evaluation (Wang)
- Horse Racing Wagering Market (Gu)
Data Mining tools
Computer Science Root

- Database
- ANN: Artificial Neural Network
- MBA: Market Basket Analysis (Association Rule)
- Genetic Algorithms
- OLAP: On-Line Analytic Processing
- Link Analysis
- High Dimensional Plots
- KDD Process
- Machine Learning
- Text Mining
Statistical Root

- (Linear & Nonlinear) Regression
- Logistic Regression
- Classification
- Clustering
- Density Estimation, Bumps and Ridges
- Time Series: Trend & Spectral Estimation
- Dimension Reduction Techniques
- High Dimensional Plots
- Classical Multivariate Methods
Statistical & Computer Science Mixture

- Data Visualization
- Dimension Reduction
  - Multi-dimensional Scaling
  - Local Linear Embedding
  - Principal Component & Principal Manifolds
  - Local Tangent Space Alignment
- Independent Component Analysis
- K-means & kd-Tree
- Support Vector Machine
Different Cultures

- Statistics/Mathematics
  - Theory
  - Formula
  - General Solution (Assumptions)
- Computer Science
  - Numerical Analysis
  - Algorithm
  - Case by Case
Why Now?

- Computer power
- The price of digital storage
- Data warehouses are being widely implemented
- Individualized marketing strategies

DeVeaux
Fundamental
- Population vs Sampling Data

Using All Data or using some data?

Data
- Sampling
- Data Squashed

Complexity
- Algorithm
- Methodology

Improved CS Mining Techniques
Statistics (Fundamental)

- Population vs Sample
- Distribution Assumption
- Parameter Estimation
- Statistical Inference
- Observations of Entire Population
- Empirical Distribution Function
- Density Estimation
- Population Description
- Forecasting

*How to analyze Population Data???
A Theory for Statistical Inference on Large Data Sets

Li, Lin and Li

Knowledge Discovery (2004)
PROPOSED PROCEDURE

Rewrite the sample as

\[ x_{11}, \ x_{12}, \ \cdots, \ x_{1\alpha_n} \quad \leftarrow \text{Block 1} \]
\[ x_{21}, \ x_{22}, \ \cdots, \ x_{2\alpha_n} \quad \leftarrow \text{Block 2} \]
\[ \vdots \quad \vdots \quad \cdots, \quad \vdots \]
\[ x_{\beta_n1}, \ x_{\beta_n2}, \ \cdots, \ x_{\beta_n\alpha_n} \quad \leftarrow \text{Block } \beta_n \]

Here \( n = \alpha_n \beta_n \)

\( \alpha_n \) to be chosen based on some criteria.

- Use the same estimator for each block,
  denoted by \( \hat{\theta}_{in} \)
- Final estimator for \( \theta(F) \)

\[ \bar{\theta} = \frac{1}{\beta_n} \sum_{i=1}^{\beta_n} \hat{\theta}_{in} \]
SAMPLING PROPERTIES

Finite Sample Properties:
- $\hat{\theta}_{in}$ unbiased
  $\implies \bar{\theta}$ unbiased
- $\hat{\theta}_{in}$ affine equivalent
  $\implies \bar{\theta}$ affine equivalent

Asymptotic Convergence
If $\alpha_n \to \infty$ and $\beta_n \to \infty$
- $\hat{\theta}_{in} \xrightarrow{P} \theta \implies \bar{\theta} \xrightarrow{P} \theta$
- $\hat{\theta}_{in} \xrightarrow{L^2} \theta \implies \bar{\theta} \xrightarrow{L^2} \theta$
- $\hat{\theta}_{in} \xrightarrow{\text{a.s.}} \theta \implies \bar{\theta} \xrightarrow{\text{a.s.}} \theta$

Share the same convergence properties.

Li, Lin & Li (2004)
Asymptotic Normality

Notation:
\[ \mu_n = E(\hat{\theta}_{in}) \]
\[ \sigma_n^2 = Var(\hat{\theta}_{in}). \]

**Condition (a)** \( \alpha_n \) is a fixed finite integer and \( \sigma_n^2 < \infty \).

**Condition (b)** \( \alpha_n \rightarrow \infty \) and \( \beta_n \rightarrow \infty \) as \( n \rightarrow \infty \), and

\[ \limsup_{n \rightarrow \infty} E \left| \frac{\hat{\theta}_{in} - \mu_n}{\sigma_n} \right|^{2+\delta} = M < \infty \]

as \( n \rightarrow \infty \) for some \( \delta > 0 \).
Theorem 1. Suppose that $x_1, \ldots, x_n$ are iid.

If either Condition (a) or (b) holds, then

$$\sqrt{\beta_n} \left( \frac{\bar{\theta} - \mu_n}{\sigma_n} \right) \to N(0, 1)$$

in distribution as $n \to \infty$. 

Li, Lin & Li (2004)
Nonparametric Kernel Density Estimation

Based on the subsample \( x_{i1}, \ldots, x_{i\alpha_n} \),

Kernel density estimator:

\[
\hat{f}_n(x; h) = \frac{1}{\alpha_n} \sum_{j=1}^{\alpha_n} K_h(x_{i,j} - x)
\]

where

\[ K_h(z) = \frac{1}{h} K(z/h), \]

\( K(z) \): a kernel density function

\( h \) is a selected bandwidth, controlling model complexity.

The choice of kernel function is not crucial,

The choice of bandwidth \( h \) is very important.

Li, Lin & Li (2004)
Bandwidth selection:

\[ \hat{h}_{\text{opt}} = \left( \frac{c_n}{n} \right)^{1/5} \hat{h}^* \]

\( \hat{h}^* \) = the optimal bandwidth using the data \( x_{i,1}, \ldots, x_{i,\alpha_n} \).

Rule of thumb:

\[ \hat{h}_{\text{rot}} = 1.06 \times s_{\alpha_n} \times n^{-1/5}, \]

Here \( s_{\alpha_n} \): sample std of \( x_{i,1}, \ldots, x_{i,\alpha_n} \).

The factor 1.06 is due to Gaussian kernel.
INTERNET TRAFFIC DATA

- Cleveland and Sun (2000, JASA)
- Cleverland, Lin, and Sun
  (2000, Proc. of ACM SIGMETRICS00)

Original data include three fields:
1. time of packet (in second)
2. direction of packet
3. size of the packet

size of dataset: \( \approx 400 \text{ MB} \)
Consists of 8.1 million nonzero throughputs
(pocket size per second)

\[ \alpha_n = \sqrt{n} \log \log(n) \]

Table 4. Estimated Percentiles of Internet Traffic Data

<table>
<thead>
<tr>
<th>( p )</th>
<th>( \hat{\pi}_p(10^6) )</th>
<th>( \hat{SE}(10^3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0015</td>
<td>0.1308</td>
</tr>
<tr>
<td>0.05</td>
<td>0.0219</td>
<td>1.3730</td>
</tr>
<tr>
<td>0.15</td>
<td>0.1120</td>
<td>4.2115</td>
</tr>
<tr>
<td>0.25</td>
<td>0.2372</td>
<td>7.2303</td>
</tr>
<tr>
<td>0.35</td>
<td>0.3836</td>
<td>9.5022</td>
</tr>
<tr>
<td>0.45</td>
<td>0.5415</td>
<td>10.3241</td>
</tr>
<tr>
<td>0.50</td>
<td>0.6300</td>
<td>10.2400</td>
</tr>
<tr>
<td>0.55</td>
<td>0.7228</td>
<td>9.8707</td>
</tr>
<tr>
<td>0.65</td>
<td>0.9033</td>
<td>8.1293</td>
</tr>
<tr>
<td>0.75</td>
<td>1.0476</td>
<td>5.1797</td>
</tr>
<tr>
<td>0.85</td>
<td>1.1329</td>
<td>2.3094</td>
</tr>
<tr>
<td>0.95</td>
<td>1.1787</td>
<td>0.8707</td>
</tr>
<tr>
<td>0.99</td>
<td>1.1858</td>
<td>0.1689</td>
</tr>
</tbody>
</table>
Figure 3: Plot of estimated density curve of internet traffic data

Li, Lin & Li (2004)
Low-Storagem Single-Pass Arbitrary Quantile Estimation

Liechty, Lin and McDermott

Journal of Statistics & Computing
Existing Methods

- **Minimax Trees**
  (Pearl, *Journal of Algorithm*, 1981)

- **Stochastic Approximation**

- **Remedian**
  (Rousseeuw and Bassett, JASA, 1990)

- **Histogram-type**
  (Hurley and Modarres, 1995)
Comments on Existing Methods

- The existing methods all perform reasonably well for median estimation.
- Only Tierney's Stochastic Approximation (S.A.) method is easily extensible to tail quantile estimation.
- The S.A. method's accuracy is dependent upon an initial sample taken to get a starting estimate and hence it is not an adequate method for tail quantile estimation.
Overview of Quantile Algorithm

- Store and sort the first 50 points and assign an estimated rank and weight to each point.
- Observe the next point, increment by 1 every point larger than the new point, and assign the new point an estimated rank and weight.
- Choose which of the now 51 points to drop based upon a score function that is derived from estimated ranks and weights for each point (The weights are used to penalize a point for being close to an existing point).
- Repeat this process until all data is exhausted.
- The point with estimated rank closest to the target rank is chosen as the estimate of the quantile.
Figure 1: Estimation of Ranks

Non-linear interpolation

Linear interpolation

Estimated rank

Grid point

\( n' \)

\( r_{m-1} \)

\( r_{i+1} \)

\( r_i \)

\( r_2 \)

\( 1 \)

\( x_1 \)

\( x_2 \)

\( x_i \)

\( x_{i+1} \)

\( x_{m-1} \)

\( x_m \)
Low-Storage Single-Pass Density Estimation

- **Univariate**
  - Certain (say, 20) representative quantiles
  - (Cubic Smoothing) Spline fitting to these quantiles
  - Statistical inference

- **Multivariate**
  - Convex Hulls Peeling
  - Certain (say, 20) representative convex hulls
  - Thin Plate Splines fitting to these quantiles
  - Statistical inference
Convex Hull Peeling Depth Contour
Astronomy Dataset: Near infra-red vs. blue-green
(subset of 11355 for illustration/comparison)
Astronomy Dataset: Depth Contour
Astronomy Dataset: Thin Plate Splines
A Specific Example: RFID