Statistical methodology for NVO

G. Jogesh Babu
Department of Statistics
Pennsylvania State University
University Park PA 16802, USA

Email: babu@stat.psu.edu

URL: http://www.stat.psu.edu/∼babu

Agenda

- History of Astrostatistics
- Role of Statistics in NVO Science
- Two new methods:
  - Data Streaming quantile estimation
  - Model selection
- NVO Web-based statistical interface
  - (in progress)

History of Astrostatistics

- Tycho Brahe (1571-1601) - Average of measurements to reduce errors
- Kepler (1571-1630) - Used arithmetic means to reject models of Mars' orbit - leading to discovery of the elliptical shapes of planetary orbits
- Galileo Galilei -"Dialogue of Two Chief World Systems" (1632) - LAD
- Adrien Legendre - (1805) "New methods for determining orbits of comets." - Determination of few unknowns from a large number of equations
- Gauss (1777-1855) and Laplace (1749-1827) - Least Squares Estimate

Various attempts over the last five centuries led to the discovery of LSE, in which Gauss and Laplace played major role.

- The leadership of astronomy in statistics waned during the second half of 19th century, mainly due to the rise of astrophysics (gravity, electromagnetism and quantum mechanics).
- 20th century – Few astronomers aware of modern statistics and few statisticians are familiar with astronomical problems.
- Statistics turned to applications in biological and social sciences and industry.
- Exception - Galaxy clustering studies by Jerzy Neyman and Elizabeth Scott (1950s and 1960s).

Recent cross-disciplinary links

Astrostatistical conferences

- *Errors, Bias and Uncertainty in Astronomy* (Jaschek & Murtagh 1989)
- *Statistical Challenges in Modern Astronomy* (Feigelson & Babu 1992; Babu & Feigelson 1997; Feigelson & Babu 2002)
- *Applications of Time Series Analysis in Astronomy & Meteorology* (SubbaRao 1997)

Monographs

- *Multivariate Data Analysis* (Murtagh & Heck 1987)
- *Astrostatistics* (Babu & Feigelson 1996)

Some broad Astrostatistical problems

• Heteroscedastic measurement errors with known variances are very common. Few statistical methods for these data.

• Parameter estimation is used to link data with astrophysical models. Data often subject to selection bias. Models often complicated and nonlinear.

• Location of galaxies in space and photons in detector constitute 2- or multidimensional spatial point processes.

• Time series. Stars and AGN exhibit enormous range of temporal behaviors.

• Censoring and truncation common in flux-limited surveys. Survival analysis is helpful but does not address all issues.

• Image restoration needed due to imperfect optics. Methods include Lucy-Richardson (= EM) algorithm and wavelet analysis.

• Bayesian methods.

NVO will present all of these problems, with the additional problem of handling massive datasets.

Why Statistics?

- Statistics can provide a clear mathematical foundation for NVO data analysis procedures.
- Statisticians evaluate the range of validity, rate of convergence, confidence measures for algorithms.
- Statistics provides a large, validated suite of preexisting methods which can be tapped for NVO use:
  - Multivariate analysis
  - Time series analysis
  - Survival analysis
  - Spatial analysis
  - Nonlinear modeling
  - Bayesian methods

Computational problems with massive datasets provided by NVO

Very simple statistic like sample mean can be estimated sequentially by updating the sum of all observations and the total number of those observations.

On the other hand, simple statistical measures such as median have no such easily implemented sequentially updatable representation.

Theoretically, a sample quantile can be obtained by sorting the data and taking the appropriate order statistic.

Sorting leads to problems with memory storage and CPU limits.

Quantile estimation for massive data

Existing Methods

- Minimax trees (Pearl, 1981)
  - Recursive tree structure where we alternatively take the minimum and maximum at successive levels of the tree.

  Problem: Restriction on the sample size that is dependent on the choice of parameters.

- Stochastic Approximation (Tierney, 1983, Lambert 2001)
  - Start with a pilot sample and get a preliminary estimate; then update it sequentially.

  Problem: Accuracy is dependent on the initial sample. It is not suitable for tail quantile estimation as it requires several passes through the whole dataset.

- Remedian (Rousseeuw and Bassett, 1990)
  - Recursive tree structure where median of b points at each of k levels are taken.

  Problem: Data size must be b^k

- Histogram-type (Hurley and Modarres, 1995)
  - Involves taking an initial sample and creating histogram-like bins and then sequentially updating counts.

  Problem: Accuracy is dependent on the initial sample.

Streaming Quantile Estimation

We developed a low-storage sequential algorithm for quantile estimation using estimated ranks and weights to calculate scores which determine the most attractive candidate data points to keep as the estimate of the quantile.

The algorithm:

1. Sort the first m data points. Set the estimated rank, r_i, for each data point, x_i, equal to the actual rank of the initial sample (i.e. $r_i = i$). Set the weight, w_i, for each data point to 1 (i.e. $w_i = 1$).
2. Determine the location of the next point in the data set, x_*, and increment the ranks of the points that are greater than the new point, i.e. if $x_i > x_*$, then $r_i = r_i + 1$.
3. Calculate an estimated rank for the new point r_*.
4. Assign a weight $w_* = \min(r_{i+1} - r_*, r_* - r_i)$, to the new point, where $x_i < x_* < x_{i+1}$.

5. Assign a score to all of the points in the array and to the new point,

$$s_\cdot = \left| \frac{r_\cdot - n'p}{w_\cdot} \right|,$$

where n' is the number of data points observed so far.

6. If the maximum score of the points being tracked is larger than the score for the new point, remove the point with the largest score from the tracking list and insert the new point, along with its estimated rank and weight, into the tracking list.

7. Repeat steps 2-6 until all elements of the data set have been seen.

8. The final estimate of the p^{th} population quantile is the point in the final array with the estimated rank closest to the target rank.

We extended this procedure to estimate multiple quantiles simultaneously.

The procedure is useful in the approximation of the unknown underlying cdf by fitting a cubic spline through the estimates obtained by this extension.

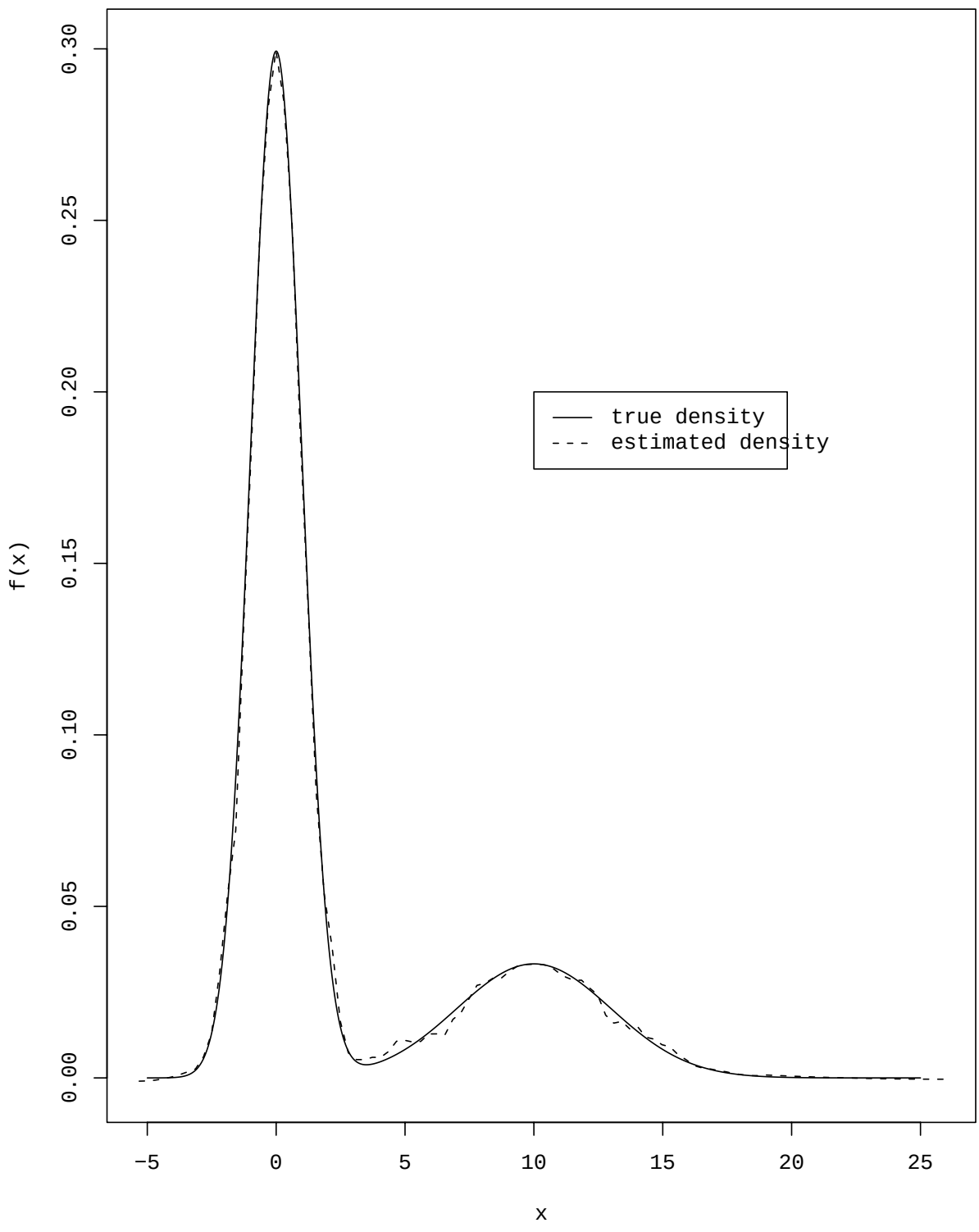
The density can be estimated by taking the derivative of this spline fit.

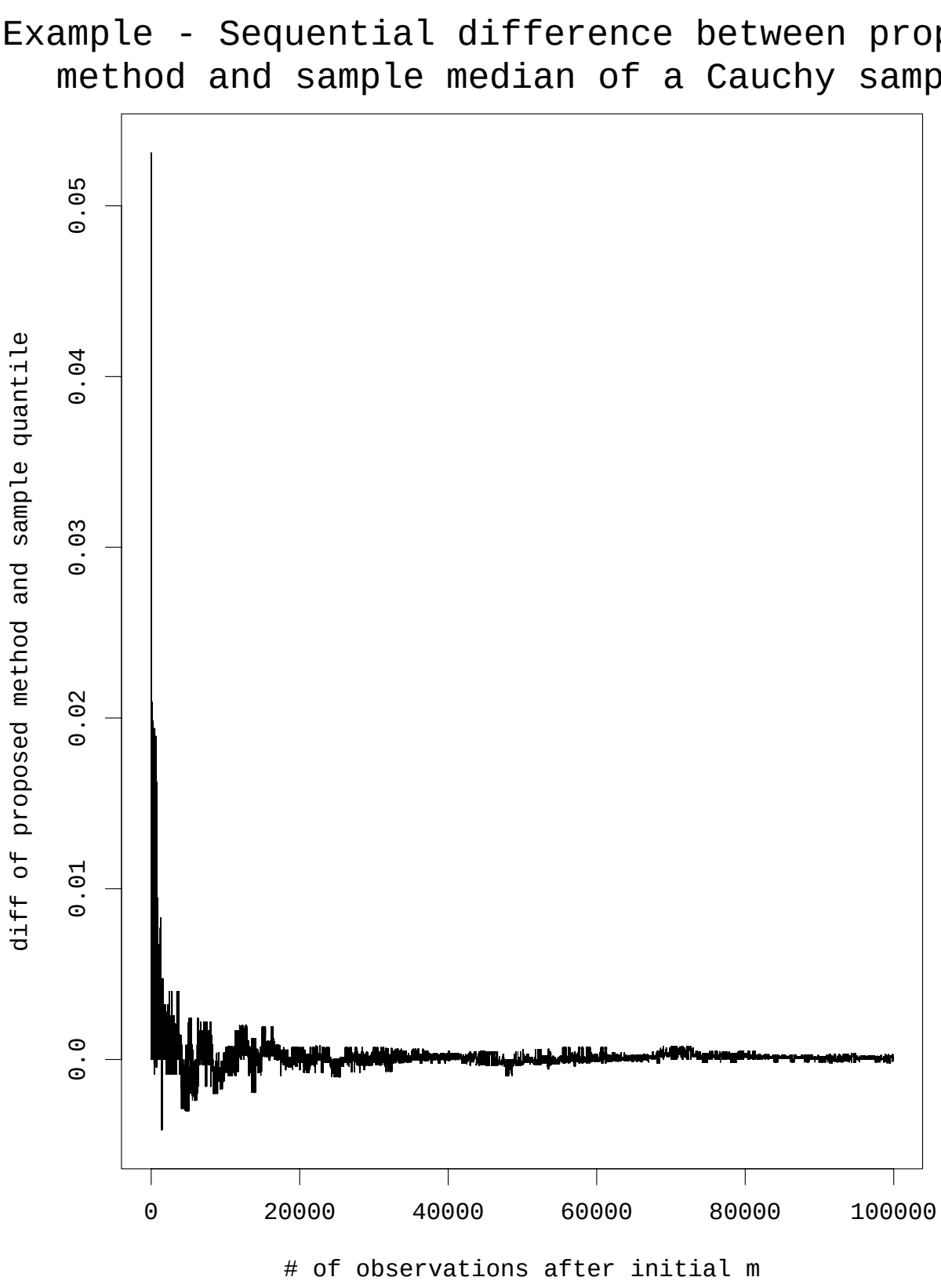
The streaming quantile estimation method is also useful in developing a 2-sample tests.

Multivariate extensions are currently under development.

Sample data set: Mixture of 2 normals with n=10,000,000.

Accurate density estimation (including tails) in $O(n)$ operations.


Example: Mixture of two normals − 75% N(0,1), 25% N(10,3)
f(x)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−5
0
5
10
15
20
25
x
true density
estimated density


Example - Sequential difference between propos
method and sample median of a Cauchy sample
diff of proposed method and sample quantile
0.0
0.01
0.02
0.03
0.04
0.05
0
20000
40000
60000
80000
100000
# of observations after initial m

Model selection

How to select among the many possible alternative models?

How to select among the many alternative curve fittings?

Does the data come from complex astrophysical model?

Popular methods include Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) – penalized likelihood.

But these methods suffer from statistical bias and other theoretical problems.

We are developing a delete-1 jackknife type method for model selection.

Advantages:

- Performs better than AIC and BIC even when the samples sizes are as small as 15 or 30. Our method picks the correct model more often than AIC or BIC.
- Does not require any new software. Use your own software to compute likelihoods and other parameters. Our method sits on top of it with only few lines of code.

No. of times the correct model is picked out of 500 simulations.

Gamma with scale parameter 5						
Normal	Data size 15			Data size 30		
s.d.	AIC	BIC	NEW	AIC	BIC	NEW
0.1	161	240	444	207	319	426
2	168	253	437	199	309	433
50	170	243	446	194	329	438
Weibull(5,0.5)						
0.1	182	227	441	191	313	420
2	165	234	432	215	291	443
50	162	226	450	198	326	440