Comments on Analyzing Data From Astronomical Surveys: Issues and Direction

Woncheol Jang

Institute of Statistics & Decision Sciences
Duke University

June 13, 2006
Outline

1. Corrnucopia of Terminology
2. Density estimation for Truncated Data with Measurement Error
3. Shrinkage and Sparsity
4. Nonparametric Bayes
This is also common in statistics: measurement error, error in variables

*We estimate that scientists are busy re-discovering America about 2/3 of time* - Simkin and Roychowdhury

Encyclopedia of Statistical Sciences, Encyclopedia of Astronomy and Astrophysics

California-Harvard Astrostatistics Group

Wikipedia
Astronomical Survey Protocols

The Nature of Survey Analysis

- Goal: estimate cosmological parameters via number density $n(r)$ and luminosity function $f(L; r)$.
- Observable: flux $F$ and direction vector $\Omega$.
- Challenges: selection bias, measurement error....
Density Deconvolution for Truncated Data

Assumption: If $m_i \in R_i(z)$, we observe

$$m_i^o = m_i + \epsilon_i$$

where $\epsilon_i$ has known density $f_\epsilon$.

In reality, astronomical measurement errors can be heteroscedastic (Akritas, 1998, Sun et al 2002)
Kernel density estimation for (Doubly) truncated data

\[ \hat{f}(x) = \int K_h(x - y) d\hat{F}_m(y) \]

where \( \hat{F}_m \) is the Lynden-Bell-Woodroffe estimator (for one sided truncation) or Efron-Petrosian estimator (for doubly truncated).

- The Efron-Petrosian estimator does not have a closed form.
- Semiparametric approach (Schafer, 2006)
Deconvolving kernel density estimator (Stefanski and Carroll, 1990)

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} K^e_h(x - m^o_j; h)$$

where $K^e_h(u; h) = K^e(u/h; h)/h$,

$$K^Z(u; h) = (2\pi)^{-1} \int e^{-itu} \{\varphi_K(t)/\varphi_{f_e}(t/h)\} dt$$

Issues: convergence rate, optimal bandwidth selection

For heteroscedastic measurement error, see Sun et al (2002)
Deconvolving kernel density estimator for (doubly) truncated data

\[ \hat{f}(x) = \int K_h^\epsilon(x - y; h) d\hat{F}_m(y) \]

where \(\hat{F}_m\) is the Lynden-Bell-Woodroofe estimator (for one sided truncation) or Efron-Petrosian estimator (for doubly truncated data) and \(K_h^\epsilon(u; h) = K^\epsilon(u/h; h)/h\)

Sun and Wang (2006) use a similar approach for biased censoring data.

It is desirable to correct the measurement error and truncation simultaneously.

Existing methods are computationally intensive.
Shrinkage plays an important role in modern statistical inferences.

Asymptotic equivalence between nonparametric function estimation and *infinite dimensional Normal means model*.

Reverse Stein Effect (Perlman and Chaudhuri, 2005): one must choose a shrinkage point with reliable knowledge about the underlying value of the parameter to be estimated.

Generalization of shrinkage estimator, modulation estimator (Beran and Dümbgen, 1998).
Sparsity

- Dimension of data grows with the number of data points
- Achieving sparsity is one of the key issues in developing relevant theory and methods
- Sparsity plays a central role in LASSO, basis pursuit and Wavelet thresholding
Is it necessary that Bayes procedures have frequentist coverage? (Berger, 2006; Wasserman, 2006)

Addressing uncertainty is a key issue in statistical inferences

Providing a measure of uncertainty in nonparametric (Bayes/Frequentist) methods is a hard problem