

**Opportunities for Modern Statistical
Methods
in Network Measurements**

David Donoho, Paul Barford, Mark Crovella, Supratik
Bhattacharya

Modern Statistics

- Huge Datasets
- Huge Numbers of Parameters
- Hidden Sparsity
- Regularization
- Multiplicity, False Discovery
- Bayesian Computations

Areas of Opportunity

- Incomplete Data
- Indirect Data
- Empirical Bayes

Incomplete Data

- Censored Observations
- Missing Covariates

Censored Observations

- Observe TCP/IP Flows during interval $[0, T]$
- Don't observe beginnings before $t = 0$
- Don't observe endings after $t = T$
- Problem: infer distribution of true durations

Similar Problems

- Napoleon's Army
- Clinical Trials
- Astronomy

Methodology

- Kaplan-Meier Estimator
- EM Algorithm
- Stochastic Imputation

Missing Covariates

- Know time delay along various paths
- Know AS paths traversed
- Don't know details inside AS
- Problem: infer delay caused by different AS's

Similar Problems

- PGE knows electricity used in house by time of day
- Don't know details of
 - Number of residents
 - Number and kind of appliances
- Problem: infer these

Methodology

- Parametrize effect of unobserved components
- EM Algorithm to estimate parameters
- Stochastic Imputation to estimate parameters

Recommended Strategy in Incomplete Data

- Say clearly what would be an ideal complete measurement
- Explicitly Model transition from ideal to real measurements
- Demand either maximum likelihood or Bayes
- Apply EM, SI, or related ideas

Indirect Data

- Example: Tomography
- Example: Wicksell's problem
- Example: MAGSAT I,II

Noise Poses Fundamental Limits

- Optics: blurry version of image: $I(x) = \int K(x, u)f(u)du$
- Math analysis gives formula: $f = L[I]$
- What prevents recovery of f ?
- Ans: noise – $Y(x) = I(x) + Z(x)$; can have $L[Y]$ blowing up.

Recommended Strategy in Indirect Data

Regularization

- Forget recovering original object
- Recover regularized (smooth) version
- Far less noise-sensitive
- Far more accurate

Regularization in Internet

- I estimate delays along $100K$ paths in Internet
- I use empirical data to optimize routes
- What is my overall behavior?
- Probably much worse than when I started
- This has nothing to do with routing instability, convergence

Empirical Bayes Methods

- Estimating the number of unseen species
- Forecasting Baseball stats

Forecasting Baseball Stats

- Background: Since 1941, no player with final season BA \geq .400
- Midseason: Suppose Top player in MLB has BA = .400
- What is best guess of his final score?
- Answer: \ll .400.

Many Normal Means

- $Y_i \sim N(\mu_i, \sigma^2)$
- What is good estimate of (μ_i)
- Traditional: $\mu_i = Y_i$.
- Always Better: $\hat{\mu}_i = \bar{Y} + (Y_i - \bar{Y}) \cdot (\hat{\tau}^2 - \sigma^2) / \hat{\tau}^2$

Estimating number of unseen species

- Observe N_1 unique species
- Observe N_2 doubletons
- ...
- Problem: N_0 (how many I didn't see)

Vivid Example: a new poem of Shakespeare is discovered; how many new words will it contain?

Simple Example

- There are N total bins equally likely
- I sample M of these uniformly at random.
- N_k = number of bins I hit k times
- $N_k \approx N \cdot \text{Poisson}_k(\lambda)$, $\lambda = M/N$
- $E(N_0) = N \cdot e^{-\lambda}$, $E(N_1) = N \cdot \lambda e^{-\lambda}$, $E(N_2) = N \cdot \lambda^2/2 \cdot e^{-\lambda}$
- $E(N_0) = 2E(N_1)^2/E(N_2)$

Conclusions

- Think about Incompleteness
- Indirectness
- Regularization