InfoUSA, NETS, and Truth Discovery

Jonathan Skaza

University of Michigan
Biostatistics for Social Impact

October 25, 2017

Outline

Introduction

- Our overarching research objective is to link features of the built-environment to human health
  - e.g., Fast food chains (FFC) within 2 mi radius of households
  - To complete our objective, we need an accurate estimate of the FFC (or the quantity of FFC) within this radius
- Two databases containing environmental features
  - *InfoUSA*: Built from over 4,000 phone directories and over 350 new business sources
  - *NETS*: Uses Dun and Bradstreet (D&B) data to create a time series database of establishment information
- Third source (e.g., *OpenStreetMap*) useful as a tie-breaker

Data Structure

Option 1: Categorical

ID	DB1	DB2	DB3
1	McDonald's, Arby's, Arby's	McDonald's, Arby's	McDonald's, Subway, Arby's
2	Subway, Burger King	Subway, Burger King	Subway, Burger King
3	Wendy's	Wendy's, McDonald's	Subway

Option 2: Numerical

ID	DB1	DB2	DB3
1	3	2	3
2	2	2	2
3	1	2	1

Data Structure Considerations

- Business names do not always match among databases
  - McDonald's vs. MCDONALDS OF MAIN STREET
  - Issue magnified with big data
  - Potential to use text similarity methods (e.g., Levenshtein distance)
- With numerical data, the businesses that comprise the count may differ among databases
  - (McDonald's, Arby's, Arby's) $\neq$ (McDonald's, Subway, Arby's)

Approaches to Truth Discovery

- Hierarchical Bayes [Cheng et al., 2014]
- Truth Finding Models [Zhao et al., 2012, Zhao and Han, 2012]

Hierarchical Bayes for Truth Discovery

Intuition

- Multiple (potentially conflicting) estimates of # of FFC within 2 mi radius
- The estimates can be combined to produce an overall, aggregate estimate

Model from [Cheng et al., 2014]

- Data from multiple surveys taken over multiple years on the same quantity of interest

$$
\begin{aligned}
y_{it}|\theta_0, \theta_t, \sigma^2_{e^*} &\sim N(\theta_t, \sigma^2_{it}) \\
\theta_t &= \theta_{t-1} + e^*_t \\
e^*_t|\sigma^2_{e^*} &\sim \text{truncated } N(0, \sigma^2_{e*})
\end{aligned}
$$

- y_{it}, σ^2_{it} known
- With our count data, we will likely want to utilize the Poisson distribution in this type of model
- Random walk model applicable to our data

Truth Finding Models

Intuition

- Conflicting information on the same entities among sources
  - Which information is correct?
  - Which sources are trustworthy?
- Some heuristics
  - There is usually only one true fact
  - The true fact is likely to appear similar among sources
  - False facts from different sources are less likely to be similar
  - A source that provides mostly true facts for many objects will likely provide true facts for other objects
- Some sources are generally more reliable than others
  - Good model of source quality is critical for truth discovery
- Given knowledge of which sources are trustworthy, we can down-weigh claims from unreliable sources
- Set of claims consistent with the overall consensus may yield information about source quality

Motivating Example from [Zhao et al., 2012]

Entity (Movie)	Attribute (Cast)	Source
Dark Knight	Christian Bale	IMDB
Dark Knight	Heath Ledger	IMDB
Dark Knight	Morgan Freeman	IMDB
Dark Knight	Christian Bale	Netflix
Dark Knight	Christian Bale	BadSource.com
Dark Knight	Heath Ledger	BadSource.com
Dark Knight	Daniel Craig	BadSource.com

Could filter false claim via majority voting. However, doing so would also erroneously detect Freeman. We may vary the threshold to 1/3 to recognize Feeman, but then Craig would be treated as true.
Solution: Model two-sided source quality. BadSource.com makes many false claims, Netflix omits true cast data.

Two-Sided Source Quality

Confusion matrix for source s

	$t = True$	$t = False$
$o = True$	TP_s	FP_s
$o = False$	FN_s	TN_s

- Precision: $\frac{TP_s}{TP_s+FP_s}$
- Accuracy: $\frac{TP_s+TN_s}{TP_s+FP_s+TN_s+FN_s}$
- Sensitivity: $\frac{TP_s}{TP_s+FN_s}$
- Specificity: $\frac{TN_s}{FP_s+TN_s}$

Evaluating Source Quality in Our Motivating Example

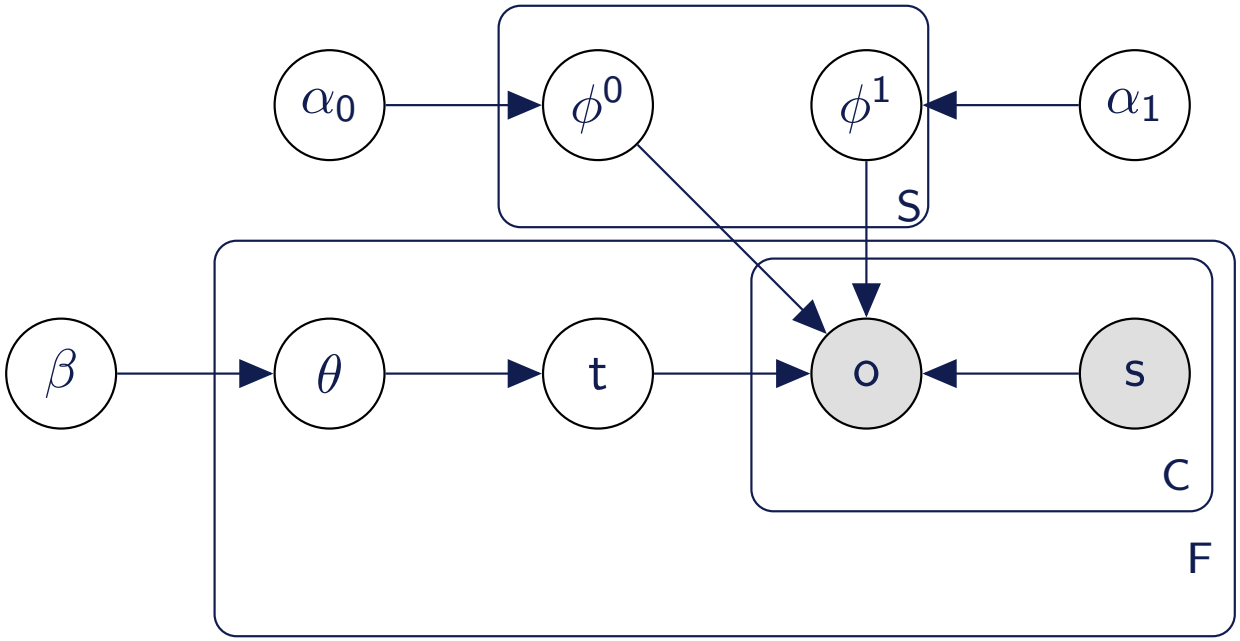
Measure	IMDB	Netlix	BadSource.com
TP	3	1	2
FP	0	0	1
FN	0	2	1
TN	1	1	0
Precision	1	1	2/3
Accuracy	1	1/2	1/2
Sensitivity	1	1/3	2/3
Specificity	1	1	0

The key contribution of [Zhao et al., 2012] is the modeling of two-sided source quality using sensitivity and specificity as two independent quality measures

Latent Truth Model (LTM) [Zhao et al., 2012]

- Treat truth and quality as latent random variables
- Given truth information, infer source quality, re-infer truth based on updated source quality
- Claims produced by high quality sources are more likely to be correct and sources that produce more correct claims are more likely to be high quality

Graphical Model of LTM



Model Details

- For each $k \in S$: generate its FPR (1-specificity), ϕ_k^0
  - $\phi_k^0 \sim Beta(\alpha_{0,1}, \alpha_{0,0})$
  - $\alpha_{0,1}$ is prior FP while $\alpha_{0,0}$ is prior TN
- For each $k \in S$: generate its sensitivity, ϕ_k^1
  - $\phi_k^1 \sim Beta(\alpha_{1,1}, \alpha_{1,0})$
  - $\alpha_{1,1}$ is prior TP while $\alpha_{1,0}$ is prior FN
- For each $f \in F$: generate its prior truth probability, θ_f
  - $\theta_f \sim Beta(\beta_1, \beta_0)$
  - β_1 is the prior true count while β_0 is the prior false count
- For each $f \in F$: generate the truth label, t_f
  - $t_f \sim Bernoulli(\theta_f)$
- For each $c \in C_f$: generate o_c
  - If $t_f = 0$, then $o_c \sim Bernoulli(\phi_{s_c}^0)$
  - If $t_f = 1$, then $o_c \sim Bernoulli(\phi_{s_c}^1)$

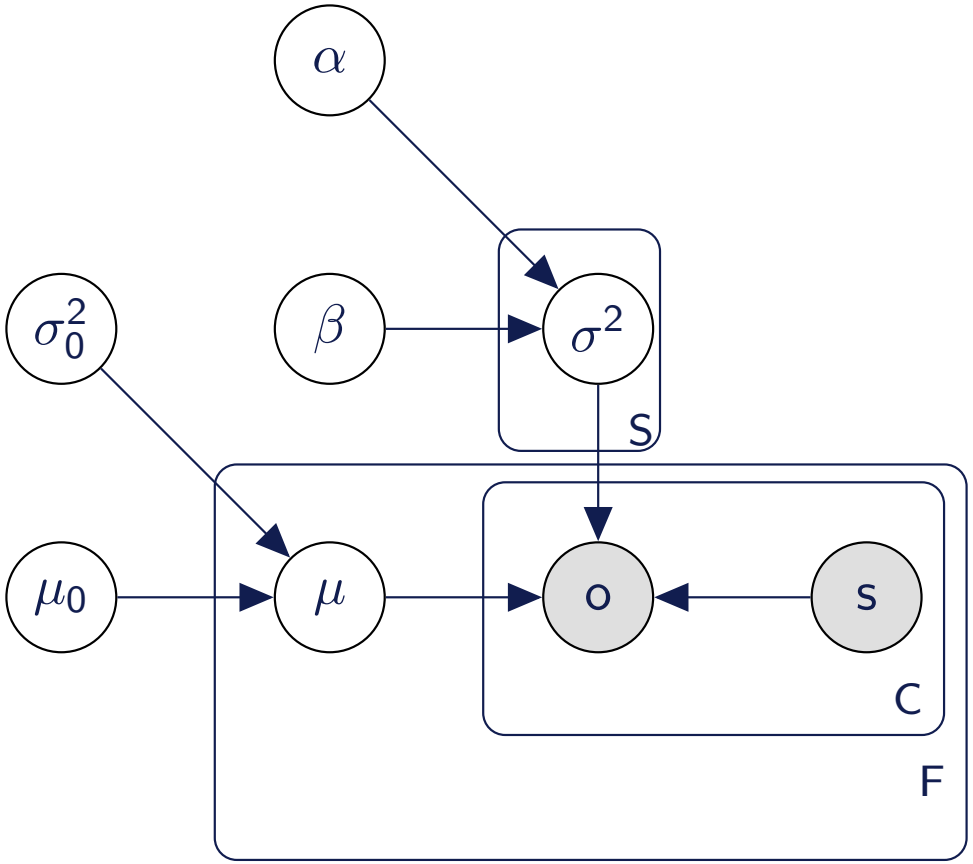
Model Considerations

- Predictions obtained using collapsed Gibbs sampling
- Maximum a posterior (MAP) estimates of source quality
- Naming conventions make data formulation a challenge

Gaussian Truth Model (GTM) [Zhao and Han, 2012]

- Truth model for numerical data
- If we have three claims—25, 140, and 150—the truth is probabaly closer to 140 and 150 than 25
- Not concerned about exact prediction as in the categorical case
- Outliers become a concern

Graphical Model of GTM



Parameter	Interpretation	Distribution
σ_s^2	Source Quality	$Inv - Gamma(\alpha, \beta)$
μ_e	Truth	$N(\mu_0, \sigma_0^2)$
o_c	Claimed Value	$N(\mu_e, \sigma_{s_c}^2)$

References


Cheng, Y., Chakraborty, A., and Datta, G. (2014).
Hierarchical Bayesian Methods for Combining Surveys.
JSM 2014 - Survey Research Methods Section.


Zhao, B. and Han, J. (2012).
A Probabilistic Model for Estimating Real-valued Truth from Conflicting Sources.
QDB.


Zhao, B., Rubinstein, B., Gemmell, J., and Han, J. (2012).
A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration.
VLDB.