Performance Bounds and Parallel Algorithms for Bayesian Data Linkage

Carnegie Mellon University

Duke University

Matt Barnes, Beka Steorts, Willie Neiswanger

Overview

- Data linkage involves identifying duplicate records in large, noisy databases.
- We investigate methods that treat data linkage as a clustering task, where each record is associated with an unobserved latent entity.
- The goal is to infer the latent entities and assignments of records to entities.
- We provide theoretical results bounding the performance of the record linkage task under this model.
- We also show preliminary work developing parallel MCMC algorithms for inference in this model, given large distributed datasets.

Introduction

- Data linkage (also known as record linkage, de-duplication, or entity resolution) involves identifying duplicate records in large, noisy databases.
- In this work, we aim to accomplish two things:
  - (1) Investigate the theoretical properties of Bayesian data linkage models and to bound aspects of the performance of these models.
  - (2) Apply parallel methods for scalable inference in data-distributed settings to these models.
- To carry out (1), we leverage these Bayesian models to (i) provide an upper bound on the KL divergence between distributions over records (over all possible linkage assignments), and use it to (ii) provide a lower bound on the probability of incorrectly estimating a record's assignment to a latent entity.
  - Note that (ii) provides a negative result concerning the intrinsic hardness of data linkage under this generative process.
- To carry out (2), we investigate how to apply recently developed embarrassingly parallel MCMC methods to this data linkage setting.

Bayesian Data Linkage Model

- **Notation** for the Bayesian data linkage model:

  - Data, denoted X, is comprised of multiple lists of records.
    k total lists
    i^{th} list has n_i records, j^{th} record has p categories
    M_l denotes number of categories in l^{th} field.
    X_{ijl} denotes l^{th} field of j^{th} record of i^{th} list.

  - Latent "deduplicated" entities, denoted Y.
    N latent entities.
    Each record corresponds to one of these latent entities.
    Assume, WLOG, that $N = \sum_{i=1}^{k} n_i$.
    $Y_{j'l}$ denotes l^{th} field of j'^{th} latent entity.

  - Linkage structure, denoted Λ, assigns latent entites to records.
    Λ_{ij} denotes assignment of j^{th} record in i^{th} list.

  - Distortion indicator, denoted z.
    $z_{ij\ell} = I(X_{ij\ell} \neq Y_{\Lambda_{ij}\ell})$, where I is an indicator function.

- **Generative process** for the Bayesian data linkage **categorical model**:

$$X_{ij\ell} \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}, \boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \begin{cases} \delta_{Y_{\Lambda_{ij}\ell}} & \text{if } z_{ij\ell} = 0 \\ \text{Multinomial}(1, \boldsymbol{\theta}_\ell) & \text{if } z_{ij\ell} = 1 \end{cases}$$

$$z_{ij\ell} \overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_\ell)$$

$$Y_{j'\ell} \mid \boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \text{Multinomial}(1, \boldsymbol{\theta}_\ell)$$

$$\boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_\ell) \quad \text{and} \quad \beta_\ell \overset{\text{ind}}{\sim} \text{Beta}(a_\ell, b_\ell) \quad \text{and} \quad \pi(\boldsymbol{\Lambda}) \propto 1,$$

- **Generative process** for the Bayesian data linkage **string model**:

- For string fields, the distorition distribution is now defined to be:

$$P(X_{ij\ell} = w \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}) = \frac{\alpha_\ell(w)\ \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}{\sum_{w \in S_\ell} \alpha_\ell(w)\ \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}$$

- We can then write the generative process as

$$X_{ij\ell} \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell} \overset{\text{ind}}{\sim} \begin{cases} \delta(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1 \text{ and } \ell \leq p_s \\ G_\ell & \text{if } z_{ij\ell} = 1 \text{ and } \ell > p_s \end{cases}$$

$$Y_{j'\ell} \overset{\text{ind}}{\sim} G_\ell$$

$$z_{ij\ell} \mid \beta_{i\ell} \overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell})$$

$$\beta_{i\ell} \overset{\text{ind}}{\sim} \text{Beta}(a, b) \quad \text{and} \quad \Lambda_{ij} \overset{\text{ind}}{\sim} \text{DiscreteUniform}(1, \ldots, N),$$

Performance Bounds

- These bounds aim to show:
  - (1) How much the distribution over the observed records can change given changes in linkage structure.
  - (2) The probability of incorrectly estimating a record's assignment to a latent entity.
- In this work, we show an upper bound, given in terms of KL divergence between distributions, for (1), and a lower bound for (2).

Statement of Performance Bounds

- Here we state the performance bounds for the Bayesian data linkage model.
- Theorem 1 is for the **categorical model**, and Theorem 2 is for the **string model**.
- let $\mathcal{P}_{ij}$ denote the set of distributions over the record X_{ij}, given the collection of all possible linkage assignments Λ_{ij}, i.e.

$$\mathcal{P}_{ij} = \{f(X \mid \boldsymbol{Y}, \Lambda_{ij}, \boldsymbol{\theta}, \boldsymbol{\beta}) : \forall \Lambda_{ij} \in \{1, \ldots, N\}.\}$$

- Let P, Q be any two distributions from $\mathcal{P}_{ij}$.

Theorem 1 *This result finds an upper bound on the KL divergence and a lower bound for the probability that model 1 gets the linkage structure incorrect. Let* $\gamma = \max_{\Lambda_{ij} \neq \Lambda'_{ij}} 2 \sum_{ij\ell} I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell})(1 - \beta_\ell) \ln\left\{\frac{1}{\min_m \theta_{\ell m} \beta_\ell}\right\}$.

i) The KL divergence is bounded above by γ. *That is,* $D_X(P||Q) \leq \gamma \ \ \forall P, Q \in \mathcal{P}$.

ii) The minimum probability of getting a latent entity wrong is $Pr(\hat{\Lambda}_{ij} \neq \Lambda_{ij}) \geq 1 - \frac{\gamma + \ln 2}{\ln r}, \ \ \forall i, j$

Theorem 2 *Assume data* $\boldsymbol{X}$, *and distributions* $P, Q \in \mathcal{P}$ *defined in section 3. Assume that two distinct linkage structures, denoted by* $Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell}$.

i) There is an upper bound on the Kullback-Leibler divergence between any $P, Q \in \mathcal{P}$ *given by* κ, *that is* $D_X(P||Q) \leq \kappa$.

ii) $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \frac{\kappa + \ln 2}{\ln r}$, *where*

$$\kappa = \max \Lambda_{ij} \neq \Lambda'_{ij} \Bigg[2 \sum_{\ell} (1 - \beta_\ell) I(Y_{\Lambda_{ij}\ell} \neq Y'_{\Lambda_{ij}\ell}) + \sum_{\ell m} I(Y_{\Lambda_{ij}\ell} \neq Y'_{\Lambda_{ij}\ell}) \left(1 - e^{-cd(Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell})}\right) E[e^{-cd(m, Y_{\Lambda_{ij}\ell})}] \Bigg] \ln\{(\min Q)^{-1}\}$$

and $r + 1$ *is the cardinality of* $\mathcal{P}$.

Empirical Results on Simulated Data

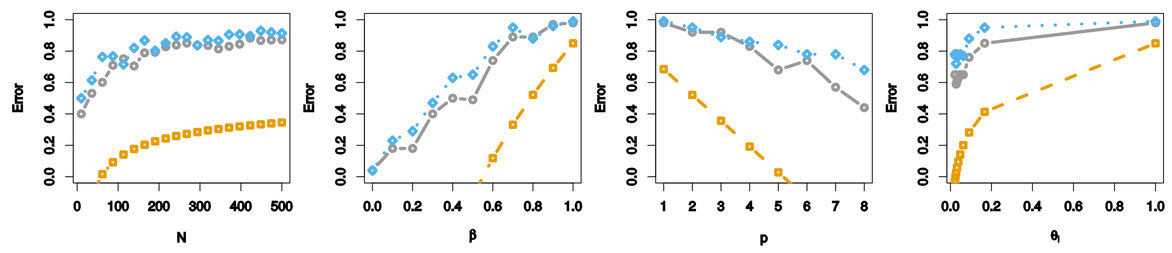

Figure 1: Theorem 1 (gold squares) holds on simulated categorical records compared to exact sampling (grey circles) and Gibbs sampler (blue triangles).

Future Work: Parallel Inference Algorithms

- **Background:**
  - We use MCMC (e.g. Gibbs sampling) for inference in this model.
  - There exist scalable MCMC algorithms for large datasets, which can process subsets of data in parallel, and then combine inferred results.
  - In particular, we hope to infer the model on each of the k lists in parallel, and then combine these results to yield a global result given the full data.

- **Goal:**
  - We hope to use **Embarrassingly Parallel MCMC**.
  - This method would carry out the following steps:
  - (1) Perform data linkage separately on each list i.
  - (2) Send inference results (i.e. samples of the latent entities Y and linkage structure Λ) , to a master machine.
  - (3) Compute a global inference: a single, global, posterior distribution over latent entities Y and linkage structure Λ), by sampling from the product of local posteriors, via existing algorithms.