Linkage Cluster Paper Draft

Brian Kundinger

Related Work

Sadinle (2017) proposed a Bayesian method for bipartite record linkage that provides a foundation for much of our proposed method. In order to ensure one-to-one matching, his Gibbs sampler conditions on the linkage status of all other records when making each linkage assignment. While this elegantly ensures transitivity, the probability distribution of the linkage vector $\mathbf{Z}$ is somewhat complicated, and does not easily allow for inclusion of additional structure among records. Our method however, motivated by recent decouple/recouple methods, initially relaxes the one-to-one matching assumption in order to incorporate linkage information differing by subgroups.

There has been some work done in the literature on conducting record linkage in situations where feild accuracy differs by subgroups of the data, but these are hampered by technical issues that make them difficult to use in practice. Larson (2005) proposed Bayesian heirarchical record linkage model that explicitly modelled the differences in parameters across blocks of the data. In relying on blocking however, he assumes that the blocking variable is error-free, when in practice such a procedure could result in many missed linkages. In contrast, our proposed method induces high probabilities of matching within a linkage cluster while still allowing the possibility to match across clusters (when the clustering field is recorded with error. Hu et al (2015) proposed a method that used a logistic regression submodel within a broader record linkage procedure to account for "expected changes" in the record of the underlying entity based on subgroup membership. This method however specified for use in temporal settings when the true attributes of the record change with time, and it is unclear how such a method would be implemented when records are collected syncronously with error, or when time information is unavailable. Additionally, the method is semi-supervised, requiring training data to implement the logistic regression, and is that not well suited for record linkage in practice. In contrast, the proposed method is fully unsupervised and is shown to not require extensive tuning of prior distributions, allowing for ease of implementation in a variety of settings.

Notation and Assumptions

Our notation and assumptions closely follow that of Sadinle (2017). Denote two files as $\mathbf{X}_A$ and $\mathbf{X}_B$, with n_A and n_B records respectively, and with records indexed as $i \in \{1, \dots, n_A\}$ in $\mathbf{X}_A$ and $j \in \{1, \dots, n_B\}$ in $\mathbf{X}_B$. Without loss of generality, label the files such that $n_A \geq n_B$. We also assume there are no duplicates within files, only across. For each record pair under consideration, we generate a comparison vector $\gamma_{ij} = \{\gamma_{ij}^1, \dots, \gamma_{ij}^F\}$, where F is the number of fields used in the linkage and each takes γ_{ij}^f takes on a value $l \in \{1, \dots, L_f\}$ indicating level agreement between the two records on a specified field.

Following the Fellegi Sunter framework, we define $m^{fl} := P(\gamma_{ij}^{f} = l | Z_j = i)$ to be the probability of observing agreement level l in field f for records i and j given that the records are a match, and similarly define $u^{fl} := P(\gamma_{ij}^{f} = l | Z_j \neq i)$, for non-matches. We denote λ to be the (marginal) probability that some record $j \in \mathbf{X}_B$ has a match in $\mathbf{X}_A$.

Lastly, we adopt from Sadinle 2017 the matching labelling $\mathbf{Z} = (Z_1, \ldots, Z_{n_A})$, defined as

$$Z_j = \begin{cases} i, & \text{if records } i \in \mathbf{X}_A \text{ and } j \in \mathbf{X}_B \text{ refer to the same entity;} \\ n_A + 1, & \text{if record } j \in \mathbf{X}_B \text{ does not have a match in file} \mathbf{X}_B; \end{cases}$$

Wherever possible, we reserve superscripts for denoting field, level, and linkage cluster, and reserve subscripts for record indices. For example, $\mathbf{m}^f = (m^{f1}, \ldots, m^{fL_f})$ is the probability distribution governing field f for matching records, and $\mathbf{m}_{ij} = \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left(m^{fl}\right)^{\mathbf{1}_{\gamma_{ij}^{f}=l}} = P(\gamma_{ij} | Z_j = i)$ is product of the relevant of the appropriate $\mathbf{m}$ parameters for record pair (i, j). We hope that these conventions avoid overloaded notation in the likelihood and subsequent derivations.

Model Specification

Additionally, our model differs from that of Sadinle 2017 through its explicit dependence on a beta random variable λ that models the rate of matching across records. Sadinle marginalizes over such a random variable in his derivations of the "beta prior for bipartite matching," but here we provide derivations without marginalizing in order to specify differing rates of matching for different linkage clusters.

Prior Distributions and Likelihood

For fields $f \in \{1, \ldots, F\}$ and levels $l \in \{1, \ldots, L_f\}$ we adopt the following likelihood and prior distributions. Note that likelihood is related to that of Sadinle (2017), but includes dependence on the λ random variable. It also makes it explict that units of observation under this model are the n_B records in $\mathbf{X}_B$, a crucial distinction when compared to the likelihood over the $n_A \times n_B$ record pairs in the original Fellegi Sunter model.

$$P(\Gamma | \mathbf{Z}, \mathbf{m}, \mathbf{u}, \lambda) = \prod_{j=1}^{n_B} \left[\lambda^{\mathbf{1}_{z_j < n_A + 1}} (1 - \lambda)^{\mathbf{1}_{z_j = n_A + 1}} \prod_{i=1}^{n_A} \mathbf{m}_{ij}^{\mathbf{1}_{z_j = i}} \mathbf{u}_{ij}^{\mathbf{1}_{z_j \neq i}} \right]$$

$$\mathbf{m^f} \sim \text{Dirichlet}(\alpha^{f1}, \ldots, \alpha^{fL_f})$$

$$\mathbf{u^f} \sim \text{Dirichlet}(\beta^{f1}, \ldots, \beta^{fL_f})$$

$$Z_j | \lambda = \begin{cases} \frac{1}{n_A} \lambda & z_j \leq n_A; \\ 1 - \lambda & z_j = n_A + 1 \end{cases}$$

$$\lambda \sim \text{Beta}(\alpha_\lambda, \beta_\lambda)$$

The prior for Z_j has equal probability of matching to all records $i \in \mathbf{X}_A$, and non-matching probability governed by λ. Observe two reasonable choices for $p(Z, \lambda)$:

If $\lambda \sim \text{Beta}(1, \frac{1}{n_A})$, then the prior for Z, integrated over λ, becomes

$$= \begin{cases} 1 & z_j \leq n_A; \\ 1 & z_j = n_A + 1 \end{cases}$$

Thus this is a uniform prior on the labellings of Z. A second useful prior would be $\lambda \sim \text{Beta}(1, 1)$, which yields

$$= \begin{cases} 1 & z_j \leq n_A; \\ n_A & z_j = n_A + 1 \end{cases}$$

Here, a match and a non-match are equaly likely, and among the matches, all records are likely to be a match. There is little difference between the the two priors in practice, as the data should overpower the prior.

Posterior Sampling

The $\mathbf{m}$ and $\mathbf{u}$ parameters are updated through standard multinomial-dirichlet mechanics. Thus we have

$$\mathbf{m}^{\mathbf{f}}|\mathbf{Z}, \Gamma \sim \text{Dirichlet}(\alpha^{f1}(\mathbf{Z}), \ldots, \alpha^{fL_f}(\mathbf{Z}))$$
$$\mathbf{u}^{\mathbf{f}}|\mathbf{Z}, \Gamma \sim \text{Dirichlet}(\beta^{f1}(\mathbf{Z}), \ldots, \beta^{fL_f}(\mathbf{Z}))$$

where $\alpha_{fl}(\mathbf{Z}) = \sum_{i,j} I_{obs}(\gamma_{ij}^f)\mathbf{1}_{\gamma_{ij}^f = l}\mathbf{1}_{z_j = i}$ and $\beta_{fl}(\mathbf{Z}) = \sum_{i,j} I_{obs}(\gamma_{ij}^f)\mathbf{1}_{\gamma_{ij}^f = l}\mathbf{1}_{z_j \neq i}$.

Since λ provides the probability that a record in $\mathbf{X}_B$ has match in $\mathbf{X}_A$, it has a full conditional distribution

$$\lambda|\mathbf{Z}^{(\mathbf{s})} \sim \text{Beta}(D + \alpha_\lambda, n_B - D + \beta_\lambda)$$

where $D = \sum_1^{n_B}(Z_j < n_A + 1)$ is the number of records identified as matches in the previous Gibbs iteration.

Lastly, we sample each Z_j according to these posterior distributions. To do so, note that as a function of Z_j, the likelihood is a discrete distibution with probabilities proportional to

$$p(\Gamma_{.j}|Z_j) \propto \mathbf{1}_{z_j < n_A + 1}\lambda \prod_{i=1}^{n_A} \mathbf{m}_{ij}^{\mathbf{1}_{z_j = i}} \mathbf{u}_{ij}^{\mathbf{1}_{z_j \neq i}} + \mathbf{1}_{z_j = n_A + 1}(1 - \lambda) \prod_{i=1}^{n_A} \mathbf{u}_{ij}$$
$$\propto \mathbf{1}_{z_j < n_A + 1}\lambda \prod_{i=1}^{n_A} \left(\frac{\mathbf{m}_{ij}}{\mathbf{u}_{ij}}\right)^{\mathbf{1}_{z_j = i}} + \mathbf{1}_{z_j = n_A + 1}(1 - \lambda)$$

$$\propto \lambda \left[\sum_{i=1}^{n_A} w_{ij}\mathbf{1}_{z_j=i}\right] + (1-\lambda)\mathbf{1}_{z_j=n_A+1}$$

where $w_{ij} = \frac{\mathbf{m}_{ij}}{\mathbf{u}_{ij}} = \frac{P(\gamma_{ij}|Z_j=i)}{P(\gamma_{ij}|Z_j\neq i)}$. The interested reader should note that these are precisely the likelihood ratios used in the Fellegi-Sunter model to classify matches and non-matches, and we therefore refer to w_{ij} as the *Fellegi Sunter weights*.

The full conditional for Z_j is thus given by

$$p(Z_j^{(s+1)}|\mathbf{m},\mathbf{u},\Gamma,\lambda,\mathbf{Z^{(s)}}) \propto p(\Gamma_{.j}|\mathbf{m},\mathbf{u},Z_j,\lambda)\, \overbrace{p(\mathbf{m},\mathbf{u}|\mathbf{Z^{(s)}})}^{\text{proportionality}}\, p(Z_j|\lambda)\;\; \overbrace{p(\lambda|\mathbf{Z^{(s)}})}^{\text{proportionality}}$$

$$\propto \left(\sum_{i=1}^{n_A} w_{ij}\mathbf{1}_{z_j=i} + \mathbf{1}_{z_j=n_A+1}\right)\left(\lambda\sum_{i=1}^{n_A}\frac{1}{n_A}\mathbf{1}_{z_j=i} + (1-\lambda)\mathbf{1}_{z_j=n_A+1}\right)$$

$$= \frac{\lambda}{n_A}\sum_{i=1}^{n_A} w_{ij}\mathbf{1}_{z_j=i} + (1-\lambda)\mathbf{1}_{z_j=n_A+1}$$

$$\implies Z_j^{(s+1)}|\mathbf{m},\mathbf{u},\Gamma,\lambda \propto \begin{cases} \frac{\lambda}{n_A}w_{ij} & z_j \leq n_A; \\ 1-\lambda & z_j = n_A+1 \end{cases}$$

Here, one should note that if we choose to integrate over the posterior of λ and rearrange terms, we produce the following sampling mechanism:

$$Z_j^{(s+1)}|\mathbf{m},\mathbf{u},\mathbf{Z^{(s)}} \propto \begin{cases} w_{ij} & z_j \leq n_A; \\ n_A\frac{n_B-D+\beta_\lambda}{D+\alpha_\lambda} & z_j = n_A+1 \end{cases}$$

The relationship between this sampling mechanism and that of Sadinle 2017 warrants discussion.

Bayes Estimate

Sadinle 2017

Decouple/Recouple

One difficulty in record linkage is conducting linkage while taking appropriate dependencies into account. In the context of this paper in which we assume that each file is recorded without duplicates, this means our linkage procedure should respect one-to-one matching. The posterior sampling of $\mathbf{Z}$ outlined above resolves half of this issue; since we only sample one potential match for record j in each Gibbs iteration, it is impossible to match j to two records at once. This in itself is an improvement over the classic Fellegi-Sunter method that labels record pairs as matches or non matches independently of all other pairs. However, our posterior sampling of $\mathbf{Z}$ does allow for the possibility that $Z_j = i = Z_{j'}$ for some $j' \neq j$, which cleary violates the assumption that $\mathbf{X}_B$ is recorded without duplicates. Sadinle 2017 resolves this issue by adapting the set of candidate matches for every Z_j, but we forgo this rigidity and instead "decouple" the linkage vector $\mathbf{Z}$ for the computational advantantages of sampling Z_j independently. In our formulation, such illegal matches occasionally occur within individual Gibbs iterations due to simple randomness, but they do not occur consistently enough between the same records to measurably influence the eventual Bayes estimate of the linkage structure.

However, if we want to ensure the theoretic purity of each MCMC sample of the linkage structure, we can "recouple" $\mathbf{Z}$ by simply resolving these illegal matchings. Explictly, define clusters $R(i) := \{j | Z_j = i\}$, and define all clusters such that $|R(i)| > 1$ as *resolution clusters*. Within each cluster, resample the candidate matches according to their Fellegi-Sunter weights. We show this step to be computationally inexpensive and to be a reasonable replication of more organic one-to-one matching. This procedure would be most useful in cases where there are many "near matches" across files, but in most cases, we find this this step to be unnecessary.

Linkage cluster model

The primary contribution of this paper is the establishment of linkage clusters, which allow the reliability parameters $\mathbf{m}^f$, the discernment parameters $\mathbf{u}^f$, and also the linkage probability parameter λ to vary according to certain attributes in the record pairs. We shall see that this set up upweights linkage within clusters and downweights linkage across clusters, thus using attribute information in a way not yet explored in the literture.

Linkage cluster assignment occurs at both at the level of the records and of the record pairs. Assign each record in each file an appropriate linkage cluster $S_k \in \{S_1, \ldots, S_k\}$, denoting the membership record j in the k^{th} cluster by $j \in S_k$. We assign the record pair comparison vector γ_{ij} to a cluster only when both records belong to that cluster; otherwise we assign it to S_0. For example

$$i \in S_1, j \in S_1 \implies \gamma_{ij} \in S_1$$

$$i \in S_1, j \in S_2 \implies \gamma_{ij} \in S_0$$

Cluster assignments at the record pair level are necessary for the updating the $\mathbf{m}$ and $\mathbf{u}$ parameters. We denote the reliability and discernment parameters assoicated with each field for each cluster as $\mathbf{m}^{fk}$ and $\mathbf{u}^{fk}$ respectively. To improve readability, we modify the definition of $\mathbf{m}_{ij}$ to incorporate the linkage cluster assignment. Explicitly,

$$\mathbf{m}_{ij} = \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left[\left(m^{flk} \right)^{\mathbf{1}_{\gamma_{ij} \in S_k}} \left(m^{fl0} \right)^{\mathbf{1}_{\gamma_{ij} \in S_0}} \right]^{\mathbf{1}_{\gamma_{ij}^{f} = l}}$$

with an analogous definition for $\mathbf{u}_{ij}$.

Additionally, define the vector of linkage probabilities $\Lambda_k = (\lambda_{k1}, \lambda_{k2}, \lambda_{k3})$, where $\lambda_{k1} =$

$P(z_{j \in S_k} \in S_k)$ is the probability that a record in $\mathbf{X}_{B_k}$ has a match within its cluster in $\mathbf{X}_{A_k}$, $\lambda_{k2} = P(z_{j \in S_k} \in S_{k'})$ is the probability that such a record $\mathbf{X}_{B_k}$ has a match outside of its cluster $\mathbf{X}_{A_{k'}}$ for some $k' \neq k$, and $\lambda_{k3} = P(z_{j \in S_k} = n_A + 1)$ is the probability of a non-match. This is a multinomial distribution where we necessarily have $\lambda_{k3} = 1 - \lambda_{k1} - \lambda_{k2}$. These parameters use linkage cluster assignment at the record level.

(THIS WOULD GO IN APPENDIX) Our new model specification becomes

$$P(\Gamma | \mathbf{Z}, \mathbf{m}, \mathbf{u}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{j \in \mathbf{X}_B^k} \left[\lambda_{k1}^{\mathbf{1}_{z_j \in S_k}} \lambda_{k2}^{\mathbf{1}_{z_j \in S_{k'}}} \lambda_{k3}^{\mathbf{1}_{z_j = n_A + 1}} \prod_{i=1}^{n_A} \mathbf{m}_{ij}^{\mathbf{1}_{z_j = i}} \mathbf{u}_{ij}^{\mathbf{1}_{z_j \neq i}} \right]$$

$$\mathbf{m^{fk}} \sim \text{Dirichlet}(\alpha^{f1k}, \ldots, \alpha^{fL_f k})$$

$$\mathbf{u^{fk}} \sim \text{Dirichlet}(\beta^{f1k}, \ldots, \beta^{fL_f k})$$

$$Z_{j \in S_k} | \Lambda_k = \begin{cases} \frac{1}{n_{A_k}} \lambda_{k1} & z_j \in S_k; \\ \frac{1}{n_A - n_{A_k}} \lambda_{k2} & z_j \in S_{k'}; \\ \lambda_{k3} & z_j = n_A + 1 \end{cases}$$

$$\Lambda \sim \text{Dirichlet}(\Lambda_0)$$

where $\Lambda_0 = (\Lambda_{k1}, \Lambda_{k2}, \Lambda_{k3})$.

THIS WOULD GO IN APPENDIX: Through similar derivations, this yields full conditionals: The **m** and **u** parameters are updated through standard multinomial-dirichlet mechanics. Thus we have

$$\mathbf{m^{fk}} | \mathbf{Z}, \Gamma \sim \text{Dirichlet}(\alpha^{f1k}(\mathbf{Z}), \ldots, \alpha^{fL_f k}(\mathbf{Z}))$$

$$\mathbf{u^{fk}} | \mathbf{Z}, \Gamma \sim \text{Dirichlet}(\beta^{f1k}(\mathbf{Z}), \ldots, \beta^{fL_f k}(\mathbf{Z}))$$

where

$$\alpha_{flk}(\mathbf{Z}) = \sum_{i,j,k} \mathbf{1}_{obs(\gamma_{ij}^f)} \mathbf{1}_{\gamma_{ij}^f = l} \mathbf{1}_{z_j = i} \mathbf{1}_{\gamma_{ij} \in S_k}$$

$$\text{and } \beta_{flk}(\mathbf{Z}) = \sum_{i,j,k} \mathbf{1}_{obs(\gamma_{ij}^f)} \mathbf{1}_{\gamma_{ij}^f = l} \mathbf{1}_{z_j \neq i} \mathbf{1}_{\gamma_{ij}^f = l}$$

We also have straightforward posterior distritutions for the linkage probability vector Λ_k:

$$\Lambda_k | \mathbf{Z} \sim \text{Dirichlet}(\mathbf{D_k} + \Lambda_0)$$

where $\mathbf{D}_k = (D_{k1}, D_{k2}, D_{k3})$ with

$$D_{k1} = \sum_{j \in \mathbf{X}_{B_k}} \mathbf{1}_{z_j \in S_k}$$

$$D_{k2} = \sum_{j \in \mathbf{X}_{B_k}} \mathbf{1}_{z_j \in S_{k'}}$$

$$\text{and } D_{k3} = \sum_{j \in \mathbf{X}_{B_k}} \mathbf{1}_{z_j = n_A + 1}$$

Our full conditional for Z_j follows a familiar form:

$$Z_{j \in S_k}^{(s+1)} | \mathbf{m}^k, \mathbf{u}^k, \Gamma, \Lambda_k = \begin{cases} \frac{\lambda_{k1}}{n_{A_k}} w_{ij} & z_j \in S_k; \\ \frac{\lambda_{k2}}{n_A - n_{A_k}} w_{ij} & z_j \in S_{k'}; \\ \lambda_{k3} & z_j = n_A + 1 \end{cases}$$

And after integrating over Λ_k and rearraning terms,

$$Z_{j \in S_k}^{(s+1)} | \mathbf{m}^k, \mathbf{u}^k, \Gamma, = \begin{cases} \frac{D_{k1} + \Lambda_{01}}{n_{A_k}(n_B - D_{k1} - D_{k2} + \Lambda_{03})} w_{ij} & z_j \in S_k; \\ \frac{D_{k2} + \Lambda_{02}}{(n_A - n_{A_k})(n_B - D_{k1} - D_{k2} + \Lambda_{03})} w_{ij} & z_j \in S_{k'}; \\ 1 & z_j = n_A + 1 \end{cases}$$