Base Model and New Linkage Cluster Model Specification

Brian Kundinger

Notation and Assumptions

Our notation and assumptions closely follow that of Sadinle (2017). Denote two files as A and B, with n_A and n_B records respectively, and with records indexed as $i \in \{1, \ldots, n_A\}$ in A and $j \in \{1, \ldots, n_B\}$ in 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 $\boldsymbol{\gamma}_{ij} = \{\gamma_{ij}^1, \ldots, \gamma_{ij}^F\}$, where F is the number of fields used in the linkage and each γ_{ij}^f takes on a value $l \in \{1, \ldots, L_f\}$ indicating level agreement between the two records on a specified field.

To indicate matching status, we adopt the *linkage structure parameter* $\mathbf{Z} = (Z_1, \ldots, Z_{n_B})$ from Sadinle 2017, defined as

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

This provides more memory efficient storage for the linkage information than a $n_A \times n_B$ sparse matrix of indicators.

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 $j \in B$ has a match in A.

Wherever possible, we reserve superscripts for denoting field and level, while reserving 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 $\mathbf{m}_{ij} = \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left(m^{fl}\right)^{\mathbf{1}_{\gamma_{ij}^f = l}} = P(\boldsymbol{\gamma}_{ij} | Z_j = i)$ 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

(This will be revised to talk more expliticly about sampling Z_j independently)

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.

$$P(\Gamma|\mathbf{Z}, \mathbf{m}, \mathbf{u}, \lambda) = \prod_{j=1}^{n_B} \prod_{i=1}^{n_A} \mathbf{m}_{ij}^{\mathbf{1}_{z_j=i}} \mathbf{u}_{ij}^{\mathbf{1}_{z_j \neq i}}$$

$$\begin{aligned}
\mathbf{m^f} &\sim \text{Dirichlet}(\alpha^{f1}, \ldots, \alpha^{fL_f}) \\
\mathbf{u^f} &\sim \text{Dirichlet}(\beta^{f1}, \ldots, \beta^{fL_f}) \\
Z_j|\lambda &= \begin{cases} \frac{1}{n_A}\lambda & z_j \leq n_A; \\ 1-\lambda & z_j = n_A + 1 \end{cases}
\end{aligned}$$

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

The prior for Z_j has equal probability of matching to all records $i \in A$, and non-matching probability governed by λ. Therefore a $\lambda \sim \text{Beta}(1, 1)$ corresponds to a prior belief that nonmatches and matches are equally likely, and a $\lambda \sim \text{Beta}(1, \frac{1}{n_A})$ prior corresponds to a uniform prior on the labelling of $\mathbf{Z}$.

Posterior Sampling

We work with the following factorization of the joint distribution:

$$p(\Gamma, \mathbf{Z}, \mathbf{m}, \mathbf{u}, \lambda) = p(\Gamma|\mathbf{Z}, \mathbf{m}, \mathbf{u}) p(\mathbf{Z}|\lambda) p(\mathbf{m}, \mathbf{u}) p(\lambda)$$

This factorization leads to following Gibbs Sampler:

Sample $\mathbf{m}^{(s+1)}$ $\mathbf{u}^{(s+1)}|\Gamma, \mathbf{Z}^{(s)}$: The $\mathbf{m}$ and $\mathbf{u}$ parameters are updated through standard multinomial-dirichlet mechanics. Thus we have

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

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

Sample $\lambda^{(s+1)}|\mathbf{Z}^{(s)}$: As a function of λ, the linkage structure parameter $\mathbf{Z}$ is sequence of successes (when $z_j < n_A + 1$) and failures (when $z_j = n_A + 1$), and therefore $p(\mathbf{Z}|\lambda) =$

$\mathcal{L}(\lambda|\mathbf{Z})$ is determined only by the number of dupliates $D = \sum_{i=1}^{n_B} \mathbf{1}_{z_j < n_A+1}$ encoded by $\mathbf{Z}$. Thus we have

$$\begin{aligned} p(\lambda|\mathbf{Z}) &\propto p(\mathbf{Z}|\lambda)p(\lambda) \\ &\propto \lambda^D(1-\lambda)^{n_B-D}\lambda^{\alpha_\lambda-1}(1-\lambda)^{\beta_\lambda-1} \\ &\propto \lambda^{D+\alpha_\lambda-1}(1-\lambda)^{n_B-D+\beta_\lambda-1} \\ \implies \lambda^{(s+1)}|\mathbf{Z}^{(s+1)} &\sim \text{Beta}(D+\alpha_\lambda, n_B-D+\beta_\lambda) \end{aligned}$$

Sample $\mathbf{Z}^{(s+1)}|\Gamma, \mathbf{m}^{(s+1)}, \mathbf{u}^{(s+1)}, \lambda^{(s+1)}$: Because we sample Z_j independently of all other $Z_{j'}$, we use only the full conditional for an individual Z_j. Let $\Gamma_{.j}$ denote the set of n_A comparison vectors with $j \in B$, and note that as a function of Z_j, the likelihood $p(\Gamma_{.j}|Z_j, \mathbf{m}, \mathbf{u}) = \mathcal{L}(Z_j|\Gamma_{.j}, \mathbf{m}, \mathbf{u})$ is a discrete distribution with probalities proportional to

$$\begin{aligned} p(\Gamma_{.j}|Z_j = z_j, \mathbf{m}, \mathbf{u}) &\propto \prod_{i=1}^{n_A} \mathbf{m}_{ij}^{\mathbf{1}_{z_j=i}} \mathbf{u}_{ij}^{\mathbf{1}_{z_j\neq i}} \\ &\propto \prod_{i=1}^{n_A} \left(\frac{\mathbf{m}_{ij}}{\mathbf{u}_{ij}}\right)^{\mathbf{1}_{z_j=i}} \qquad \text{By dividing through by } \prod_{i=1}^{n_A} \mathbf{u}_{ij} \\ &= \begin{cases} w_{ij} & z_j \leq n_A; \\ 1 & z_j = n_A + 1 \end{cases} \end{aligned}$$

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*.

With the likelihood in this form, we can derive the full conditional

$$p(Z_j|\Gamma_{.j}, \mathbf{m}, \mathbf{u}, \lambda) \propto p(\Gamma_{.j}|Z_j, \mathbf{m}, \mathbf{u})P(Z_j|\lambda)$$

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

In order to make fair comparisons against the Sadinle 2017 model, we integrate over the posterior of λ and rearrange terms to produce the final full conditoinal:

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

Decouple/Recouple

(This will be revised; I'm not sure if I really believe the "Recouple" step anymore! I'll need look into it more on larger datasets)

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 sets $R(i) := \{j|Z_j = i\}$, and define all sets such that $|R(i)| > 1$ as *resolution sets*. Within each set, 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.

Bayes Estimate

Same as Sadinle.

Efficient Computation

Adopting some from Enamorado 2019, some steps my own. very promising, stay tuned!

Proposed Linkage Cluster Model

(Still thinking through best notation and terminology)

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.

Partition files A and B into *record clusters* $\{A_1, \ldots, A_k\}$ and $\{B_1, \ldots, B_k\}$ according to some attribute in the data. We assign the record pair (i, j) to a *linkage cluster* $S_k \in \{S_1, \ldots, S_K\}$ only when both records belong the corresponding record clusters; otherwise we assign it to S_0. For example

$$i \in A_1, j \in B_1 \implies \gamma_{ij} \in S_1$$
$$i \in A_1, j \in B_2 \implies \gamma_{ij} \in S_0$$

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,

$$\mathbf{m}_{ij} = \prod_{f=1}^{F} \prod_{l=1}^{L_f} \prod_{k=0}^{K} \left[\left(m^{flk} \right)^{\mathbf{1}_{\gamma_{ij} \in S_k}} \right]^{\mathbf{1}_{\gamma_{ij}^f = l}}$$

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 B_k has a match within its cluster in A_k, $\lambda_{k2} = P(z_{j \in S_k} \in S_{k'})$ is the probability that such a record B_k has a match outside of its cluster, in $\mathbf{X}_{A_{k'}}$ 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.

Our new model specification becomes

$$P(\Gamma | \mathbf{Z}, \mathbf{m}, \mathbf{u}, \lambda) = \prod_{j=1}^{n_B} \prod_{i=1}^{n_A} \mathbf{m}_{ij}^{\mathbf{1}_{z_j = i}} \mathbf{u}_{ij}^{\mathbf{1}_{z_j \neq i}}$$

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

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

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

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

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

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

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

where

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

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

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

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

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

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

Our full conditional for Z_j follows a familiar form:

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

And after integrating over Λ_k and rearranging terms,

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