

Base Model and New Linkage Cluster Model

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 $i \in \{1, \dots, n_A\}$ in A and $j \in \{1, \dots, n_B\}$ 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 $\gamma_{ij} = \{\gamma_{ij}^1, \dots, \gamma_{ij}^F\}$, where F is the number of fields used in the linkage and each γ_{ij}^f takes on a value $l \in \{1, \dots, L_f\}$ indicating level agreement between the two records on a specified field.

To indicate matching status, we adopt the *linkage structure parameter* $\mathbf{Z} = (Z_1, \dots, Z_{n_A})$ 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, $\mathbf{m}^f = (m^{f1}, \dots, m^{fL_f})$ is the probability distribution governing field f for matching records, and

$\mathbf{m}_{ij} = \prod_{f=1}^F \prod_{l=1}^{L_f} (m^{fl})^{\mathbf{1}_{\gamma_{ij}^f=l}} = P(\gamma_{ij} | Z_j = i)$ 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

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 $f \in \{1, \dots, F\}$ and levels $l \in \{1, \dots, L_f\}$ we adopt the following likelihood and prior distributions.

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

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

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

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

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

The prior for Z_j has equal probability of matching to all records $i \in A$, 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

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

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

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

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 $\Gamma_{\cdot j}$ 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 probabilities proportional to

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

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

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

Here, we integrate over the posterior of λ and rearrange terms to produce the final full conditional:

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