1 MULTIPLE MATCH

We provide updated notation to allow us to describe one record in B having multiple matches in A. Let Z_j be a set containing the indices for all of the records in A that are a match with record B_j, and let $Z = \{Z_j | j = 1, \ldots, n_B\}$ denote the collection of such sets for all records in B. Let $|Z_j| = \sum_{k=1}^{\infty} I(Z_{j,k} > 0)$ denote the number of records in A that are linked to B_j. We use $Z_j = \emptyset$ to denote when B_j has no match in A.

We can allow each record in B to match to multiple records in A through a Dirichlet process prior. Define a vector of probabilities $\boldsymbol{\pi} = (\pi_0, \ldots)$ where π_k is the probability that some record in B has exactly k matches in A. In implementation, we model each π_k as a product of conditional probabilities: let η_k be the probability that some record in B has at least k matches, given that it has at least $k-1$ matches. This gives us the stick breaking representation

$$\pi_k = (1 - \eta_{k+1}) \prod_{c=1}^{k} \eta_c, \tag{1}$$

where η_k are independent random variables from a Beta$(\alpha_\eta, \beta_\eta)$ distribution.

Note that when considering possible values that Z_j can take, the order of its elements are irrelevant for record linkage. That is, $Z_j = (i, i')$ and $Z_j = (i', i)$ both communicate that record B_j is matched to records A_i and $A_{i'}$. Let $\sigma(q)$ denote all possible orderings of the elements of q. We define the equivalence relation $q' \equiv q$ if and only if $q' \in \sigma(q)$, and note that $|\sigma(q)| = |q|!$.

Similar to fabl, we adopt a prior specification on Z so that each matching Z_j of length $|Z_j| = k$ is equally likely. We can write this prior in two equivalent ways. We can write the prior in terms of a particular vector q as

$$p(Z_j = q | \boldsymbol{\pi}) = \frac{(n_A - |q|)!}{n_A!} \pi_{|q|}. \tag{2}$$

Equivalently, we can write the prior in terms of the equivalence class $\sigma(q)$ as

$$p(Z_j \equiv q | \boldsymbol{\pi}) = \frac{(n_A - |q|)!|q|!}{n_A!} \pi_{|q|}, \tag{3}$$

where the additional $|q|!$ accounts for the multiple orderings of q. Though the representation in (3) is more natural in the record linkage settings, the representation in (2) will provide more clear derivations in the sections that follow.

1.1 Derivation of Joint Distribution

We first derive the Gibbs sampler for a the joint distribution of the vector Z_j. Following the observation of Wortman (2019) and elaborated by Kundinger et al. (2024), when B_j does not link to any record in A (such that $|Z_j| = 0$) the contribution to the likelihood is simply a product of u parameters, which we will call c_j:

$$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_j = \emptyset) = \prod_{i=1}^{n_A} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} = c_j. \tag{4}$$

When $Z_j = q = (q_1, \ldots, q_k)$ for some $|q| > 0$, we have

$$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_j = q) = \prod_{i \in q} \prod_{f=1}^{F} \prod_{l=1}^{L_f} m_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \prod_{i \notin q} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)}. \tag{5}$$

We multiply and divide by the u parameters for the matching record pairs to obtain

$$\begin{aligned} p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_j = q) &= \prod_{i \in q} \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left(\frac{m_{fl}}{u_{fl}}\right)^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \prod_{i=1}^{n_A} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} & (6) \\ &= c_j \prod_{i \in q} w_{ij}. & (7) \end{aligned}$$

Lastly, we multiply the likelihood by the prior in (2) to obtain the posterior distribution. For $Z_j = q$ where $|q| = k$, we have

$$\begin{aligned} p\left(Z_j = q|\gamma, \boldsymbol{m}, \boldsymbol{u}, \pi\right) &= \frac{\frac{(n_A - k)!}{n_A!} \pi_k c_j \prod_{i \in q} w_{ij}}{\sum_{h \in \mathcal{Z}} \frac{(n_A - |h|)!}{n_A!} \pi_{|h|} c_j \prod_{i \in h} w_{ij}} & \text{(8a)} \\ &= \frac{\frac{(n_A - k)!}{n_A!} \pi_k \prod_{i \in q} w_{ij}}{\sum_{h \in \mathcal{Z}} \frac{(n_A - |h|)!}{n_A!} \pi_{|h|} \prod_{i \in h} w_{ij}} & \text{(8b)} \\ &\propto \frac{(n_A - k)!}{n_A!} \pi_k \prod_{i \in q} w_{ij} & \text{(8c)} \end{aligned}$$

Importantly, the constant c_j is not found in the final expression because the probability mass associated with every potential value for Z_j shares the same c_j. This does not occur due to proportionality. We emphasize that this full conditional is for one particular representation q of the equivalence class $\sigma(q)$. The full conditional for entire class of representations is given by

$$p\left(Z_j \equiv q|\gamma, \boldsymbol{m}, \boldsymbol{u}, \pi\right) \propto \frac{(n_A - k)!k!}{n_A!}\pi_k \prod_{i \in q} w_{ij}. \tag{9}$$

1.2 Sequential Sampler

Sampling this joint distribution is computationally prohibitive as the number of records in A grows. In particular, when allowing B_j to match to up to k records, there are $\sum_{c=1}^{k} \frac{n_A!}{(n_A-c)!c!}$ possible options for the set Z_j. Furthermore, if we were to sample this joint distribution directly, we would need to choose a maximum k ahead of time in order to probably enumerate these potential options. Through Gibbs sampling however, we can break this joint distribution into a sequence of more simple conditional univariate distributions. This allows for a more computationally efficient sampler, and allows us to learn k from the data, rather than set it ahead of time.

We generalize the fast beta prior from Kundinger et al. (2024) to a sequence of priors that allows for multiple matchings. When B_j has been linked to $k-1$ records, we say that the probability that B_j has a k^{th} match is η_k, and that all remaining records in A are equally likely to be linked. let $Z_{j,-k} = (Z_{j,1}, \ldots, Z_{j,k-1})$ be the set of records linked to B_j before the k^{th} matching phase. We use

$$p(Z_{j,k} = q_k|\eta_k) = \begin{cases} \frac{\eta_k}{n_A-(k-1)}, & q_k \notin N_{j,k}, \\ 1-\eta_k, & z_{j,k} = \emptyset; \end{cases} \tag{10}$$

where $N_{j,k} = [n_A] \setminus Z_{j,-k}$ is the set of records in A that are available to be matched with B_j. This sequence of priors leads to sequence of posteriors that can be used to sample arbitrarily many links for record B_j. These posteriors are given by

$$p(Z_{j,k} = q_k|Z_{j,k-1}, \eta_k, \boldsymbol{m}, \boldsymbol{u}, \gamma) \propto \begin{cases} \frac{\eta_k}{n_A-(k-1)} w_{q_k,j}, & q_k \in N_{j,k}, \\ 1-\eta_k, & q_k = \emptyset, \end{cases} \tag{11}$$

as derived in Appendix 2.1.

This sequential sampler produces an output $Z_j = q = (q_1, \dots, q_k)$ when $Z_{j,c} = q_c$ for steps $c \in \{1, \dots, k\}$, and the $k+1$ step produces $Z_{j,k+1} = \emptyset$. Observe that

$$p(Z_{j,k+1} = \emptyset | \Gamma_{.j}, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{\eta}) \prod_{c=1}^{k} p(Z_{j,c} = q_c | \Gamma_{.j}, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{\eta}) \tag{12}$$

$$\propto (1 - \eta_{k+1}) \prod_{c=1}^{k} \frac{\eta_c}{n_A - (c-1)} \prod_{c=1}^{k} w_{q_c,j} \tag{13}$$

$$= \frac{(n_A - k)!}{n_A!} (1 - \eta_{k+1}) \prod_{c=1}^{k} \eta_c \prod_{c=1}^{k} w_{q_c,j} \tag{14}$$

$$= \frac{(n_A - k)!}{n_A!} \pi_k \prod_{c=1}^{k} w_{q_c,j} \tag{15}$$

$$= p\left(Z_j = q | \gamma, \boldsymbol{m}, \boldsymbol{u}, \pi\right). \tag{16}$$

Since the output of the sequential sampler is necessarily ordered, it produces a particular represention for a set of matches as shown in (8c), rather than the full equivalence class as shown in (9).

This sequential sampler amounts to an extension of `fabl` with an iterative matching phase. In each iteration of the Gibbs sampler, we sample an initial set of links using η_1. For each record in B that was found to have a link, we remove the linked record in A from consideration, and then sample another potential link with η_2. We continue, using η_k in the k^{th} matching step, until no new links are found, at which we point the matching phase terminates. The $\boldsymbol{\eta}, \boldsymbol{m}$, and $\boldsymbol{u}$ parameters are estimated based on all of the links identified, regardless of the order in which they are sampled. Crucially, there is no need to specify a maximum number of links per record, as this estimated through the model.

REFERENCES

Kundinger, B., Reiter, J. & Steorts, R. C. (2024), 'Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)', *Bayesian Analysis (Accepted)* .

Wortman, J. P. H. (2019), Record linkage methods with applications to causal inference and election voting data, PhD thesis, Duke University.

2 APPENDIX

2.1 Derivation of Sequential Sampler

We now provide the derivation of the sequential sampler, following the argument presented in Section 1.1. Suppose B_j has been linked to k records in A. Let $Z_{j,-k} = (Z_{j,1}, \ldots, Z_{j,k-1})$ denote the vector of records already linked to B_j. When B_j has no additional link in A, the contribution to the likelihood is a product of the u parameters for all remaining records. That is,

$$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_{j,k} = \emptyset, Z_{j,-k} = q_{-k}) = \prod_{i \notin Z_{j,-k}} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} = c_{Z_{j,-k}}. \tag{17}$$

When $Z_{j,k} = q_k$ for some $q_k > 0$, we have

$$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_{j,k} = q_k, Z_{j,-k} = q_{-k}) = \prod_{f=1}^{F} \prod_{l=1}^{L_f} m_{fl}^{I(\gamma_{q_k,j}^f = l) I_{obs}(\gamma_{q_k,j}^f)} \prod_{i \notin (q_{-k}, q_k)} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \tag{18}$$

$$= \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left(\frac{m_{fl}}{u_{fl}} \right)^{I(\gamma_{q_k,j}^f = l) I_{obs}(\gamma_{q_k,j}^f)} \prod_{i \notin (Z_{j,-k}} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \tag{19}$$

$$= c_{Z_{j,-k}} w_{q_k,j} \tag{20}$$

To obtain the posterior, we multiply by the prior in (10). The posterior distribution this is given by

$$p(Z_{j,k} = q_k | Z_{j,k-1}, \eta_k, \boldsymbol{m}, \boldsymbol{u}, \gamma) = \frac{\left(\frac{\eta_k}{n_A - (k-1)} c_{Z_{j,-k}} w_{q_k,j} \right)^{I(q_k \in N_{j,k})} + \left(c_{Z_{j,-k}} (1 - \eta_k) \right)^{I(q_k = \emptyset)}}{\frac{\eta_k}{n_A - (k-1)} c_{Z_{j,-k}} \sum_{i \notin Z_{j,-k}} w_{ij} + c_{Z_{j,-k}} (1 - \eta_k)} \tag{21}$$

$$= \frac{\left(\frac{\eta_k}{n_A - (k-1)} w_{q_k,j} \right)^{I(q_k \in N_{j,k})} + (1 - \eta_k)^{I(q_k = \emptyset)}}{\frac{\eta_k}{n_A - (k-1)} \sum_{i \notin Z_{j,-k}} w_{ij} + (1 - \eta_k)} \tag{22}$$

$$\propto \begin{cases} \frac{\eta_k}{n_A - (k-1)} w_{q_k,j}, & q_k \in N_{j,k}, \\ 1 - \eta_k, & q_k = \emptyset. \end{cases} \tag{23}$$