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 all such set. Let $|Z_j| = \sum_{k=1}^{\infty} I(Z_{j,k} > 0)$ denote the number of links associated with record 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.

Similar to fabl, we adopt a prior specification on Z so that each matching Z_j of length $|Z_j| = k$ is equally likely. Formally, we have

$$\mathcal{L}(Z, \boldsymbol{m}, \boldsymbol{u} \mid \gamma) = \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left[m_{fl}^{I(i \in Z_j)} u_{fl}^{I(i \notin Z_j)} \right]^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)}, \tag{2a}$$

$$\boldsymbol{m}_f \sim \text{Dirichlet}(\alpha_{f1}, \ldots, \alpha_{fL_f}), \forall f = 1, \ldots, F, \tag{2b}$$

$$\boldsymbol{u}_f \sim \text{Dirichlet}(\beta_{f1}, \ldots, \beta_{fL_f}), \forall f = 1, \ldots, F, \tag{2c}$$

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

$$\pi_k = (1 - \eta_{k+1}) \prod_{c=1}^{k} \eta_c, \tag{2e}$$

$$\eta_k \sim \text{Beta}(\alpha_\eta, \beta_\eta). \tag{2f}$$

1.1 Derivation of Joint Distribution

We first derive the Gibbs sampler for a the joint distribution of the set 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:

$$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_j = \emptyset) = \prod_{i=1}^{n_A} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} = c_j. \tag{3}$$

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

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

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

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

$$= c_j \prod_{i \in q} w_{ij}. \tag{6}$$

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

$$p\left(Z_j = q|\gamma, \boldsymbol{m}, \boldsymbol{u}, \pi\right) = \frac{\frac{(n_A - k)!k!}{n_A!} \pi_k c_j \prod_{i \in q} w_{ij}}{\sum_{h \in \mathcal{Z}} \frac{(n_A - |h|)!|h|!}{n_A!} \pi_{|h|} c_j \prod_{i \in h} w_{ij}} \tag{7}$$

$$= \frac{\frac{(n_A - k)!k!}{n_A!} \pi_k \prod_{i \in q} w_{ij}}{\sum_{h \in \mathcal{Z}} \frac{(n_A - |h|)!|h|!}{n_A!} \pi_{|h|} \prod_{i \in h} w_{ij}} \tag{8}$$

$$\propto \frac{(n_A - k)!k!}{n_A!} \pi_k \prod_{i \in q} w_{ij} \tag{9}$$

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.

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

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

$$\eta_k \sim \text{Beta}(\alpha_\eta, \beta_\eta) \tag{11}$$

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. Through an argument analogous to that of Section 1.1 (see Appendix 2.1), we have,

$$p(Z_{j,k} = q_k|Z_{j,k-1}, \eta_k, \boldsymbol{m}, \boldsymbol{u}, \gamma) \propto \begin{cases} \frac{\eta_k}{n_A-(k+1)} w_{q_k,j}, & q_k \in N_{j,k}, \\ 1-\eta_k, & q_k = \emptyset; \end{cases} \tag{12}$$

$$\eta_k \sim \text{Beta}(\alpha_\eta + n_k(Z), \beta_\eta + n_{k-1}(Z) - n_k(Z)) \tag{13}$$

where $n_k(Z) = \sum_{j=1}^{n_B} I(|Z_j| \geq k)$ is the number of records in B that have at least k matches in A. Note that $n_0(Z) = n_B$, and that for each k, we can view $n_{k-1}(Z)$ as a number of trials, and $n_k(Z)$ as a number of successes, and thus update η_k as a straightforward Beta random variable.

While the set Z_j is unordered, the sequential sampler necessarily samples $Z_{j,k}$ in an ordered fashion. Let $\sigma(Z_j)$ denote every possible ordering of the elements of Z_j. Because

we regard each element $\sigma(Z_j)$ as equal, they must be equally likely. Therefore we have

$$\prod_{q' \in \sigma(q)} \prod_{k=1}^{|q|} p(Z_{j,k} = q'_k | \Gamma, m, u, \eta_k, Z_{j,k-1}) = |q|! \prod_{k=1}^{|q|} p(Z_{j,k} = q_k | \Gamma, m, u, \eta_k, Z_{j,k-1}). \quad (14)$$

With this established, we see that the joint posterior for the set Z_j provided in (7) is proportional to the product of sequential posteriors provided in (12). For $Z_j = q$ where $|q| = k$, we have

$$p\left(Z_j = q | \gamma, \boldsymbol{m}, \boldsymbol{u}, \pi\right) \propto \frac{(n_A - k)!k!}{n_A!} \pi_k \prod_{i \in q} w_{ij} \quad (15)$$

$$= \prod_{c=1}^{k} \frac{1}{n_A - (c+1)} (1 - \eta_{k+1}) k! \prod_{c=1}^{k} \eta_k \prod_{c=1}^{k} w_{q_c,j} \quad (16)$$

$$= (1 - \eta_{k+1}) k! \prod_{c=1}^{k} \frac{\eta_k}{n_A - (c+1)} \prod_{c=1}^{k} w_{q_c,j} \quad (17)$$

$$\propto p(Z_{j,k+1} = \emptyset | \eta_k) k! \prod_{c=1}^{k} p(Z_{j,c} = q_c | \eta_c) p(\Gamma_{\cdot j} | \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{\eta}, q_c \in Z_j) \quad (18)$$

$$\propto p(Z_{j,k+1} = \emptyset | \Gamma_{\cdot j}, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{\eta}) k! \prod_{c=1}^{k} p(Z_{j,c} = q_c | \Gamma_{\cdot j}, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{\eta}) \quad (19)$$

$$= p(Z_{j,k+1} = \emptyset | \Gamma_{\cdot j}, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{\eta}) \prod_{q' \in \sigma(q)} \prod_{c=1}^{k} p(Z_{j,c} = q'_c | \Gamma_{\cdot j}, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{\eta}) \quad (20)$$

As such, the joint distribution for any set Z_j of length k is proportional to the probability of any sequence of matching phases that attains all elements of Z_j times the probability that the sequential samplers stops at the $k+1$ step.

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,

$$p(\Gamma_{\cdot j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_{j,k} = \emptyset, Z_{j,-k} = q_{-k}) = \prod_{i \notin Z_{j,-k}} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} = c_{Z_{j,-k}}. \tag{21}$$

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

$$p(\Gamma_{\cdot j}|\boldsymbol{m}, \boldsymbol{u}, \pi, Z_{j,k} = q_k, Z_{j,-k} = q_{-k}) = \prod_{f=1}^{F} \prod_{l=1}^{L_f} m_{fl}^{I(\gamma_{q_k,j}^f = l) I_{obs}(\gamma_{q_k,j}^f)} \prod_{i \notin (q_{-k}, q_k)} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \tag{22}$$

$$= \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left(\frac{m_{fl}}{u_{fl}} \right)^{I(\gamma_{q_k,j}^f = l) I_{obs}(\gamma_{q_k,j}^f)} \prod_{i \notin (Z_{j,-k}} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \tag{23}$$

$$= c_{Z_{j,-k}} w_{q_k,j} \tag{24}$$

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

given by

$$p(Z_{j,k} = q_k | Z_{j,k-1}, \eta_k, \boldsymbol{m}, \boldsymbol{u}, \gamma) = \frac{\left(\frac{\eta_k}{n_A-(k+1)} c_{Z_{j,-k}} w_{q_k,j}\right)^{I(q_k \in N_{j,k})} + \left(c_{Z_{j,-k}}(1-\eta_k)\right)^{I(q_k=\emptyset)}}{\frac{\eta_k}{n_A-(k+1)} c_{Z_{j,-k}} \sum_{i \notin Z_{j,-k}} w_{ij} + c_{Z_{j,-k}}(1-\eta_k)} \tag{25}$$

$$= \frac{\left(\frac{\eta_k}{n_A-(k+1)} w_{q_k,j}\right)^{I(q_k \in N_{j,k})} + (1-\eta_k)^{I(q_k=\emptyset)}}{\frac{\eta_k}{n_A-(k+1)} \sum_{i \notin Z_{j,-k}} w_{ij} + (1-\eta_k)} \tag{26}$$

$$\propto \begin{cases} \frac{\eta_k}{n_A-(k+1)} w_{q_k,j}, & q_k \in N_{j,k}, \\ 1-\eta_k, & q_k = \emptyset. \end{cases} \tag{27}$$