Response to Reviewer Comments

May 11, 2023

Thank you for the thoughtful feedback. We agree that much of the notation iscumbersome, and we in particular thank the reviewers for seeing the merit in our methods despite those difficulties. We have made several broad changes to address the Associate Editor's overall critiques, and have addressed each specific comment from the reviewer.

1 AE Comments

AE.0: I must add that reading the paper still feels like a chore, in part because the notation is so heavy, which may be hard to avoid, but also, in part, because some wording choices make certain sentences hard to parse, which can certainly be improved.

> *We agree that our notation is at times cumbersome. We have made several efforts to simplify notation in this revision.*
>
> - *Instead of referring to files as $\mathbf{X}_1$ and $\mathbf{X}_2$, we are now using A and B. We find this generally more readable, and particularly simplifies the notation in Section 4.2.*
> - *At the beginning of Section 4.3, we have rewritten the likelihood function in terms of the agreement patterns and the summary statistics discussed in Section 4.1. This is the text added:*
>   *"To ground the discussion of posterior inference, we first re-express the likelihood function from (6a) in terms of the agreement patterns and summary statistics in Table 2. Let $m(p) = p(\gamma_{ij} = h_p | Z_j = i)$ be probability that records A_i and B_j form agreement pattern p given that they are a match. Similarly define $u(p)$. These probabilities are simply products of relevant $\boldsymbol{m}$ and $\boldsymbol{u}$*

> *parameters. Viewed through the perspective of patterns, the likelihood is*
>
> $$\mathcal{L}(\boldsymbol{Z}, \boldsymbol{m}, \boldsymbol{u} \mid \tilde{\boldsymbol{\Gamma}}) = \prod_{j=1}^{n_B} \prod_{p=1}^{P} \prod_{i \in r_{p_j}} m(p)^{I(Z_j=i)} u(p)^{1-I(Z_j=i)}.\text{"} \tag{1}$$
>
> *This likelihood is then used throughout the discussion of efficient posterior inference. We hope this expression clarifies the meaning of "preservation of weight (Comment 5), and strengthens the argument for $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{N}\}$ being sufficient statistics for $\boldsymbol{m}$, $\boldsymbol{u}$, and $\boldsymbol{Z}$ (Comment 4).*
>
> - *To denote when the (i, j) record pair exhibits agreement pattern p, we have removed the clunky $(i, j) \in h_p$ notation for $\gamma_{ij} = h_p$, which we feel is more direct and intuitive.*
> - *Section 4.1 introduces the sufficient statistic N_{j_p}. After careful review, we realized that by changing the order of subscripts and instead using N_{p_j}, we can replace the sum of counts across all records (previously denoted H_p), with the more intuitive N_p. Thus, we have removed all of the "H" quantities throughout Section 4. We believe this substantially improves readability.*

AE.1: Line 80 on p. 3 says: "For ease of readability, we follow the convention established by Sadinle (2017) and say "record $i \in X_1$" rather than the more compact x_{1i}." Taking this at face value implies that, in any sentence, one could swap "record $i \in X_1$" for "x_{1i}," which clearly is not true. So, as it stands, the sentence does little to improve ease of readability. In fact, if the point is that "x_{1i}" will never be used again, why is the notation introduced in the first place?

> *We added this sentence in an attempt to address a previous reviewer comment pointing out a conflation between the record itself and the index directing towards that record. We did in fact intend for "record $i \in X1$" to be mean precisely x_1i. However, we see now that introducing x_{1i} has introduced new confusion. With the simplified reference to the two datasets as A and B, that material now reads:*
>
> *"Consider a data files A and B, consisting of records A_i and B_j respectively, where $i \in \{1, \ldots n_A\}$ and $j \in \{1, \ldots n_B\}$".*

2 Reviewer Comments

R.0: For the missing data treatment now included just before Section 2.1, is the assumption truly missing at random, or missing completely at random? Or, does this distinction not matter because of the independence assumed across elements of the comparison vector?

> *I used MAR because that is what Sadinle wrote in his paper. Please review, and confirm whether its MCAR or not.*
>
> *The assumption is missing completely at random. You are correct that this is a result of the independence assumed across elements of the comparison vector. This has been clarified.*

R.1: Appendix 8.2: I appreciate the streamlining of the discussion of the proposed algorithms and think it was a good choice to move the full derivation to an appendix. However, I do not follow the re-expression of the pmf for $\Gamma_{.j}$. First, the opening square brackets are still misplaced throughout. In the second line of this derivation, the authors divide the expression by a product of the element-wise conditional probabilities of a match in the comparison vector (u), raised by an indicator that the element of the comparison vector equals a particular value. Perhaps I misunderstand, but I believe this product is not constant in $\Gamma_{.j}$ or u, and so the total expression is not proportional to the line above, as interpreted as a pmf. I think that the end effect is that the last line of the newly expressed pmf is missing a factor of $\prod_i \prod_f \prod_l u_{fl}^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)}$ regardless of the value of z_j.

> *The square brackets have been corrected, and we apologize for the oversight.*
>
> *We have provided an alternative derivation of the full conditional, separately handling the cases for $Z_j \leq n_A$ and for $Z_j = n_A + j$. We provide it here for your reference:*
>
> *We consider the likelihood in two cases. When B_j does not link to any record in A, the contribution to the likelihood is simply a product of u parameters, which we will call c_j:*
>
> $$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, Z_j = n_A + j) = \prod_{i=1}^{n_A} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} = c_j. \tag{2}$$
>
> *When $Z_j = i$ for some $i \leq n_A$, we have*
>
> $$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, Z_j = i) = \prod_{f=1}^{F} \prod_{l=1}^{L_f} m_{fl}^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} \times \prod_{i'!=i} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{i'j}^f=l)I_{obs}(\gamma_{i'j}^f)}. \tag{3}$$
>
> *We multiply and divide by the u parameters for the matching record pair to obtain*
>
> $$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, Z_j = i) = \prod_{f=1}^{F} \prod_{l=1}^{L_f} \left(\frac{m_{fl}}{u_{fl}}\right)^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} \times \prod_{i'=1}^{n_A} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{i'j}^f=l)I_{obs}(\gamma_{i'j}^f)} \tag{4}$$
>
> $$= w_{ij} \times c_j. \tag{5}$$

> *Combining these results, we have*
>
> $$p(\Gamma_{.j}|\boldsymbol{m}, \boldsymbol{u}, Z_j = z_j) = \begin{cases} w_{z_j,j} & z_j \leq n_A; \\ 1 & z_j = n_A + j, \end{cases} \tag{6}$$

R.2: Appendix 8.2: Thank you for including the details on integrating out pi from the full conditionals. They surprised me. I had assumed that the authors had integrated out pi in the prior for Z, as this is what Sadinle (2017) had done to form the "beta prior for bipartite matchings". Does the alternate approach presented in this paper provide a different algorithm than directly integrating the prior distribution? Is the presented alternative approach justified?

> *The reviewer raises important questions. The main difference between the full conditionals in Sadinle (2017) and ours is that for him, Z is a partition, while for us, it is not. This is what required him to integrate out π in the prior, and then sequentially update each component of the Z vector.*
>
> *The approach we have is justified, as indicated in the proof provided. However, under closer review, we have determined that it is unnecessary. Without integrating out π, all paramaters are updated through a standard Gibbs sampler. Furthermore, in other work we are developing, we actually need samples of π in different applications.*
>
> *We reran all case studies and simulation using samples of π in the full conditional for Z, and found no difference in results. This is unsuprsing, because with high numbers of observations, the posterior of the beta distrubution is highly concentrated around its mean.*
>
> *Therefore, we have updated the paper with full conditionals for the standard Gibbs sampler. We leave the remark about marginalization in the Appendix to highlight the similarity to Sadinle's method.*

R.3: I do not understand the statement just below Equation (10): "When j has no match in X_1, we write $(n_1 + j, j) \in h_{P+1}$" My understanding of these patterns is that they are based on observed comparison vectors without consideration of Z (matches). In the second paragraph of Section 4.3, the H notation includes the matches (Z), in notation and definition that seems to conflict with the statement just below Equation (10). This also comes into play in Equation (16)

> *You are correct that the the comparison vectors are created (and therefore, patterns are assigned) without regard to Z. This notation was created purely to be able to denote when a record was left unlinked during the Gibbs Sampler as in Equation 16*

> *in the submitted draft.*
>
> *I can see why this is confusing. Therefore, I have removed that line, and changed the notation for record B_j being unmatched to just be"otherwise".*

R.4: Third paragraph of Section 4.1: The authors claim they are computing "sufficient statistics". What exactly are these statistics sufficient for?

> *We show in the revised Section 4.3 that $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{N}\}$ is a set of statistics that are sufficient for m, u, and Z.*
>
> *I drafted a whole bunch of material to argue more rigoursly that $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{N}\}$ is set of sufficient statistics. If this argument seems weak, or is too much new material, I can also just keep what I had, and use less rigorous term "summary statistics" instead. [TODO]*

R.5: First paragraph of Section 4.2: The authors state: "Posterior calculations still attribute the appropriate weight to all records through the summary statistics..." What is meant by the term "weight"? Which records are appropriately weighted – those in X2?

> *For each $j \in \{1, \ldots, n_B\}$, I use SEI to eliminate from memory the indices of many records in A that are unlikely matches for B_j However, since I know the number of such records that exhibit each pattern p, I can still include their contribution to the likelihood through N_{p_j}.*
>
> *We have changed that sentence in the text to read "All original record pairs still contribute to the likelihood through $\mathcal{N}$, and thus we can proceed with posterior inference through the memory reduced $\tilde{\Gamma}^{SEI} = \{\mathcal{P}, \mathcal{R}^{SEI}, \mathcal{N}\}$."*
>
> *This is clearer when looking at the conditional likelihoods in Equations 15 and 17 using $\tilde{\Gamma}$ in the revised Section 4.3.*

R.6: Second paragraph of Section 4.2: "and delete those comparison vectors". Which are "those" vectors?

> *We have revised that sentence to read "We then conduct hashing, obtain the compressed $\tilde{\Gamma}^{ab}$ for later calculations, and delete the larger Γ^{ab} from memory before continuing with the next chunk of data."*

R.7: Where does $R^{SEI,cd}$ come into play in the partitioned algorithm presented in Equations (13) and (14)? I recommend that the authors either refrain from suppressing the SEI notation or further explain how the SEI algorithm has changed the quantities in these equations.

> *The written expression in (13a) is correct. We then compress $r_{p_j}^{cd}$, and store in memory $r_{p_j}^{SEI,cd}$ Equation (14a) is conducted through $r_{p_j}^{SEI,cd}$. Equations (13b-c) and (14b-c) are exact, unchanged by SEI. We have clarified this.*

R.8: Section 4.2: I appreciate the practical advice about choosing S for the SEI method. However, this choice seems arbitrary in the absence of further discussion/evidence. Given that the primary novelty of the manuscript is in methods to speed and otherwise improve computation, I am surprised that this aspect of computational innovation is presented with virtually no theoretical or empirical exploration. Presumably the SEI method has some sort of accuracy trade-off, as the authors warn that linkage results may be "distorted" if S is low. However, this trade-off is not quantified or even discussed in practical terms beyond the terse recommendation to choose S=10.

> *I'm not sure the best way to address this. Here is the where the situation arises:*
>
> *Imagine file B has one record B_1 for Brian Kundinger, and file A has 20 nearly matching records for Brian Kundinger. Perhaps Brian A Kundinger, Brian B Kundinger, Brian C Kundinger, ..., ect. All of these records share the same agreement pattern (say $p = 2$) with B_1.*
>
> *Without SEI, posterior probability would be divided evenly among all 20 near matches, such that the posterior probability would be at most $\frac{1}{20}$. If I chose $S = 2$ however and I only retained 2 representatives in r_{2_1}, I could likely have one of those records having posterior probability above .5, and I would erroneously declare it a match.*
>
> *So issues only arise in very degenerate situations. How do you think I should address this? Should I do a simulation showing degeneracy when S=1 or 2?*

R.9: Page 5, 2 sentences before equation (4): I believe the sum should be of $I(Z_j \leq n_1)$, not $I(Z_j \leq n_1 + 1)$.

> *That is correct. This has been corrected, and we apologize for the oversight.*

R.10: Equation (6a): The indices do not match the subscripts in the indicator function in each summand, or their standard meaning in table 1.

> *I do not understand this objection.*

R.11: Generally, the authors seem to arbitrarily use upper and lower case z interchangeably in function definitions.

> *When writing, I attempted to use Z when I was discussing a random quantity, and z to reference a realized value. I haven't found anything that I see wrong. I think I'm going to need help with this.*

R.12: Equations (8) and (9): Should the weights have superscript (s) (as the Zs do)?

> *When writing, I believed it was sufficient to include the superscripts on the m and u in the definition of w. However, you are correct, it is more clear to explicitly include it on the w. This has been updated.*

R.13: Gamma is in some places described as a set and in others as a matrix (particularly in sections 4.2 and 5.1).

> Γ *is a matrix comprised by comparison vectors. It is certainly not a set in the mathematical sense. All references to* Γ *as a set have been removed.*
>
> *I found one example of me saying that* Γ *is "comprised of the set of comparison vectors.' I'll triple check before resubmission.*

R.14: Section 4.3, second paragraph, the definitions of the concatenated vectors α_0 and β_0 should have final elements subscripted by L_f, where the sub-subscript is capitalized.

> *You are correct. This has been revised.*