Response to Reviewer Comments

July 6, 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 end of Section 4.1, we have rewritten the likelihood function in terms of the agreement patterns and the summary statistics. This is the text added: This likelihood is then used throughout Section 4.2 in 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 summary 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 summary 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 $x_1 i$. 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?

> *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 $\Gamma_{\cdot j}$. 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 $\Gamma_{\cdot j}$ 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 also prepared an alternative derivation of the likelihood for Z_j, separately handling the case when B_j has a match and when it does not. We believe this is more clear.*

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 unsurprising, because with high numbers of observations, the posterior of the beta distribution 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 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 4.2 that the perspective of patterns allows us to write conditional likelihoods and posterior updates for $\boldsymbol{m}$, $\boldsymbol{u}$, and $\boldsymbol{Z}$ entirely throught the statistics in $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{N}\}$. However, we cannot express the full likelihood in (14) through these statistics. Therefore, we have revised to text to say "summary statistics" instead of the more technical "sufficient statistics."*

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?

> *We have switched the order of Sections 4.2 and 4.3, so that we first discuss hashing, then posterior inference, then chunkwise computation of the comparison matrix, and then SEI. With this new ordering, and with the newlyu provided equations (15 -19) it is clearer that the contribution of* all *record pairs is recorded through the summary statistics in $\mathcal{N}$.*

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.

> *We have reordered Section 4 so that all posterior inference is presented before SEI. This makes it clear that the posterior updates in (16a), (16b), and (18) depend only*

> *on $\mathcal{N}$. SEI only affects the step shown in (19). We have provided an additional simulation study to show robustness to the SEI procedure.*

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.

> *We have provided an additional simulation study in Section 5.3 to show robustness to different choices of S.*

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.

> *This has been fixed.*

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

> *We use Z when discussing a random quantity, and z to reference a realized value. We have revised, and have found all instances of inconsistency.*

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

> *This has been fixed.*

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.*

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.*