Response Markdown

2022-08-31

Editor

Dear Dr Kundinger,

Thank you for submitting your paper "Efficient and Scalable Bipartite Matching through Fast Beta Linkage (fabl)" for possible publication in Bayesian Analysis. It has now been carefully reviewed and my decision is: Major revision required.

These are the comments I received from the Editor who has handled your paper:

The paper has been reviewed by a referee and an associate editor and I have read the manuscript independently, before looking at the reports. When I confront my notes and overall impression with the reviewers' comments, I find myself in substantial agreement with the reviewers.

As indicated by both reviewers, the work expands on previous work by Sadinle, introducing computational advances that are achieved by relaxing some requirements in the original formulation of the problem. The close connections with the previous work and the extent of the novel contributions are not adequately explained. This makes it hard for the reader to understand what is new and why the new contributions are valuable. A careful editing of the presentation is needed to address this shortcoming.

Regarding the presentation, the AE points out a number of problems with notation, typos, and various inconsistencies. Independently, I found many similar problems. The intersection between what I found and what the AE found is non-empty and so is the symmetric difference. Also, the AE found some issues that I did not find and vice versa. So, I am almost certain that an additional review would uncover more problems. Now, the results presented in the paper seem plausible as do the broad strokes of the derivations, but I must admit that I was not able to follow all the details. This, in large part, was because of the issues that I just mentioned, and I believe that most readers would find it difficult to follow the developments, as these issues are exceedingly distracting.

There are problems with the notation, which in places is not defined, in others is used before being defined, and in others yet is used inconsistently.

For example, $I_obs()$ is never defined, $n_12(Z)$ first appears on p. 5, but is not defined (in passing) until the bottom of p. 7, and it seems to become D on p. 8.

> How and when should I define I_{obs}? I probably need to lay out certain assumptions about missingness?
>
> Thank you for identifying inconsistencies with the notation $n_{12}(Z)$. We have made these corrections.

Throughout the manuscript, n_1 and n_2 are used interchangeably with n_A and n_B, sometimes in the same section, as it happens, for example, in Section 3.1.

> Thank you for addressing this oversight. We have made these corrections throughout the text.

The loss function on p. 9 is out of Sadinle's paper and it inherits the typo therein (θ_{11}, $if Z_j, \hat{Z}_j, \dots$).

> Thank you for addressing this oversight. We have corrected the loss function.

In the statement of Lemma 1 on p.13, B comes out of nowhere. (Thinking I had missed something, I went back to the previous pages only to find out that B would be then defined in the proof.)

We have updated the Lemma to include the definition of B.

Formatting of all displayed math must be checked and fixed, especially as far as punctuation is concerned, as commas seem to be missing in multiple places. For example, in Equation (1), there should be a comma after 0 and after 1, and this is by no means an isolated occurrence.

Is this true?

Confusion can also arise from lack of clarity in the exposition. Take comment 4 from the AE, for example. Whose marriage certificates and whose birth records are involved? What records are in X_1 and what records are in X_2? The reader should not be expected to go back to Newcombe's paper to clarify the issue.

Please see comment in Reviewer 2 responses.

This is not an exhaustive list, but only a set of examples meant to point out what makes the paper hard to read and how it can be improved. Many more specific points are raised in the AE's report and here are a couple of other minor issues I noticed:

p. 12, second paragraph: insert "is" after Section 3.1

Thank you for identifying this. We have made the correction.

p. 15: the first sentence of the last full paragraph is either missing a verb ("we observe"?) or it includes an extra "that."

I can't find this typo

As I said, I do not believe that we, as reviewers, were able to uncover all the problems, and I strongly encourage the authors to do their part, as they should, to improve the presentation and eliminate all typos and inconsistencies.

The AE finds the simulation studies to be incomplete and gives detailed suggestions on how those should be improved. The AE also gives important suggestions about other aspects of the manuscript that must be carefully considered and addressed. All of the AE's comments are right on the mark.

To summarize, the paper contains an interesting algorithmic contribution that can speed up calculations at the expense of relaxing some of the modeling conditions, without much adverse impact on the resulting inferences. This aspect should be emphasized in a revised presentation. The presentation should also make clearer the close connections to the relevant work by Sadinle, and all the issues that I mentioned above should be ironed out.

The authors should prepare a careful and substantive revision that remains within the editorial limit of 25 pages and that answers the various comments satisfactorily. The authors should upload the revised manuscript together with a document detailing how they addressed the reviewers' comments.

Reviewer 2

This manuscript addresses the bipartite matching problem in data linkage and proposes a new fast computation version of "Beta Linkage" to allow such matching to take place for pairs of large databases without resorting to ad hoc blocking. It is built upon the Felegi-Sunter model of record linkages, where linkages are made within strata of agreement between records across the two databases, and relies heavily on advances made by Sadinle (2017).

Overall, the paper presents a practical implementation of a modified Beta Linkage approach to matching. It demonstrates that even though the new method suffers from the limitations of not seamlessly enforcing one-to-one matching, it is feasible for large datasets with modest numbers of features on which to match.

Though the novelty of the presented methodology is primarily in the computational approach, I believe it to be a real advance for practical implementation of record linkage in modern applications.

I have no major concerns about the content or accuracy of the presentation. However, substantial improvements are warranted to confirm my impression. In addition to the recommendation made by Reviewer 1 to include more clarity of the novel contributions of this manuscript, I believe that the explanation of the method should be written more clearly, the simulation comparisons should be repeated/expanded, and various features of the simplification should be fleshed out a bit more. Finally, more practical advice would be welcome. All of these points are more specifically described in the itemized list below

Overall, the notation is, somewhat by necessity, quite extensive. However, I'm not convinced that it's entirely consistent. A glossary in the appendix may help the reader navigate the paper more easily.

I don't think we require an appendix. It's not much more notation than Sadinle or Fellegi Sunter.

Section 2, line 1: It may help to explicitly define X1 and X2 as being vectors of indices as impliedby equations (1) and (2).

I'm not sure I agree with this point

Section 2.2, beta distribution: the $n_{12}(.)$ notation needs to be defined.

We thank the reviewers for identifying this oversight, and have defined the notation.

Section 3, 4th full paragraph: My understanding of the example of matching birth records and marriage certificates (presuming monogamy) may not quite be consistent with the algorithm as presented, as I would think that one certificate (element in X1) would match multiple birth records (elements in X2) which violates the size ordering. Can the algorithm be easily adjusted for this reverse sizing?

PLEASE READ: Part of me thinks we should just remove this element of the model from the paper. This topic also begs the question of whether linkages stay the same depending on which dataset you label X1 and X2. I have no good way of answering that.

Essentially though,the answer is yes, the algorithm can easily be adapted for the reverse sizing. The issue is mostly notational.

It seems as though Sadinle chose to label the datasets X_1 and X_2 such that $n_1 > n_2$ for mostly notational reasons. He uses the notation $Z_j = n_1 + j$ when record $j \in X_1$ has no match in X_2, which works because when $n_1 > n_2$, $n_1 + j$ never refers to any record in X_2. Jody uses the simplified notation $Z_j = n_1 + 1$ to refer to unmatched records, and this works for the same reason.

However, when we relax the one-to-one constraint, or even encourage a modeller to allow multiple linkages in one dataset, we shouldn't denote the datasets according to simple notational convenience. This might require clever edits of the "Review of Prior Work" section, and other areas of the paper.

Notationally, we can use $Z_j = 0$ or $Z_j = -j$ as a way to unambiguously denote a record as nonmatching.

But again, it may be easier to just strike the paragraph about the advantages of many-to-one matchings.

Section 3, full model: $I_obs(.)$ seems to be taken from a different paper dealing with missing fields, which is not at all discussed in this manuscript. This consideration of missing fields and the accompanying discussion of missing data mechanism should be included if such data is to be considered for this manuscript. This comment also applies to the $Gamma^{obs}$ notation in the following subsection.

I can add these definitions, no problem. But is it worth including mention of missing data?

Section 3, full model: The Phi notation needs to be defined.

We thank the reviewers for identifying this oversight, and have defined the notation.

Section 3, full model: The square brackets seem to be misplaced, as the exponent includes indices of the sum contained in the brackets.

We thank the reviewers for identifying this oversight, and have corrected the expression.

Section 3.1, I found the motivational descriptions in this section to be somewhat confusing, in that they seem to mix the posterior behavior (e.g., parameters are updated though standard multinomial-Dirichlet...) and model construction (as a function of pi, Z is a series of successes...). This subsection might be easier to understand if its purpose was more bluntly described in an opening sentence or paragraph and streamlined to focus only on the full conditional distributions for a Gibbs Sampler. In addition, the mixed motivation may have led to errors in intermediate conclusions. In particular, the derivation of the probability of $Gamma_{\cdot j}$ seems to be missing a factor of $u^I_{fl}(gamma^f_{ij} = l)$. This is not important to the eventual full conditional of Z, but does not seem to be correct for the distribution of Gamma as claimed.

I'm not entirely sure how to address the first part of the response. Should I just eliminate a lot of the exposition? Also, I see no issue with $p(\Gamma_{\cdot j})$

Section 3.1, full joint posterior distribution: I was surprised to see that the contribution of the conditional distribution of Z was included using the summation-style notation rather than the binomial/multinomial-style multiplicative notation that naturally motivates the full conditional presented on the first line of page 8.

The contribution of Z in the full conditional is just the prior distribution, which we have expressed with summation notation to reflect the piecewise nature of the prior. The binomial-looking expression at the top of page 8 is the posterior distribution for π. We note that the posterior $p(\pi|\Gamma, Z, \Phi)$ in fact only depends on Z, and is thus shortened to $p(\pi|Z)$. We have updated our explanation to make this more clear.

Section 3.1, first equation on page 8: There is an extra parenthesis in the exponent of (1-pi).

We thank the reviewers for identifying this oversight, and have corrected the expression.

Section 3.1, pmf for $Gamma_{\cdot j}$ and full conditional for Z: The notation in the final equation may be more clear if you replace i with z_j, as in $w_{z_j,j}$. I believe that "$n_1 + 1$" should be "$n_1 + j$" in several places.

Ask Beka if i or z_j is more clear.

We agree that using z_j instead of i in that context is more clear. We have also changed $n_1 + 1$ to $n_1 + j$ in several instances. Thank you for bringing this to our attention.

Section 3.1, last equations on page 8: There seems to be some lack of specificity in describing these equations as full conditional distributions vs. steps in the Gibbs Sampler algorithm. It would be helpful to provide a brief justification of integrating out pi.

Should I just remove this step? I'm not sure how to justify this step, but also this step isn't actually necessary. We could just sample π and use it, as in a standard Gibbs sampler. The only reason to integrate it out is to see the similarity to Sadinle's full conditional, but its not necessary at all.

Section 4.2, SEI procedure. I'm not sure I follow the description of the SEI procedure. My understanding of SEI is that for patterns with lots of possible pairs, no single pair is likely to (should) be identified in the posterior distribution. Thus, you take a small subsample of the records of size $S < H_{j_p}$, and store these in R^{SEI} rather than the complete index of pairs in each R. Can you quantify the computational savings due to this method? Do you have advice for

how small S can be? In your simulations, how does the use of this method or choice of S affect the results?

All computational savings occur from conducting the Gibbs sampler over the P patterns instead of the n_2 records. SEI is for memory savings. I think I need help articulating this clearly.

Section 4.4, Assumptions and definitions should be included in the statement of this lemma.

Do I really need this to be a Lemma? It seems overly clunky to me, but if its convention to provide computational complexity as a lemma, thats OK with me.

Section 5.1 simulations – there is a detailed comparison for one set of simulation settings, including a binary definition of a "match". Do these results differ as these settings change?

No, the time savings would be the same. In all cases, the fabl method replaces dependency on n_2 for dependency on P, the number of patterns. To me, this is clear, I'm not sure how to clarify this in the paper

Section 5.2 and 5.3 simulations – are these a single simulated datasets? Do these same results hold over multiple simulated datasets, Or was this just a chance result for each setting?

In the caption under Figure 3, we explain that we have 100 pairs of sets of records for each level of overlap and error. We have added this description in the body of the paper as well for clarity.

Section 5.3 simulation – RR results are not presented.

I will redo simulations in order to get this

Section 6.1 last sentences. I believe the last sentence should be removed, as it is redundant.

We have deleted this last sentence. Thank you for addressing this oversight.

Section 6.2, second sentence. I believe the word "survey" is missing, as in "the initial SURVEY began. . . "

Thank you for addressing this oversight.

Section 6.2. In the NLTCS, are participants added in later cohorts, or may we assume that every member of the smaller dataset should also be included in the earlier one?

Beka, can you help answer this?

Figures 5 and 7 are not referenced in the text.

Thank you for addressing this oversight. We have included pointers to these figures in the main text.

Reviewer 3

The paper presents a computational variant of the method introduced by Sadinle (2017). Simulation results are presented and the variant is applied on a new, significantly larger data set.

A significant part of the paper is a reproduction of the paper of Sadinle. The variant is introduced in section 4. It is simple but consequential. The simulations and applications show how the variant accelerates the computation relative to the original method and reveals the trade-offs. The value of the paper resides in that computational complexity continues to be a difficult obstacle in the application of Bayesian statistics. The proposals of the paper enable the application of Sadinle's approach to an extent not possible before.

It is important to present the research for what it is: An extension of the method of Sadinle motivated by computational considerations. Here Sadinle supersedes Fellegi-Sunter in the sense while the starting point of Sadinle's research is Fellegi-sunter, the starting point of this paper is Sadinle's work. This should be made clearer, beginning in the abstract and reiterated in the conclusion.