Response to Reviewer Comments

January 27, 2024

We thank the Editor and Associate Editor for their assessment and constructive feedback.

1 Editor Comments

E.0: Anyhow, I have decided to follow the AE's recommendation and would like the authors to carefully consider the comments in the AE report and make the suggested revisions. I find the issue raised by the AE regarding the formulas in Section 4.2 particularly concerning for two reasons. First, the article cannot be accepted if it contains technical errors and inaccuracies. Second, the formulas in question were used to code up the samplers used in the simulations and data analysis examples. If the formulas are incorrect (do not make sense, in fact, as the AE points out) how can the computational results be trusted? This is a grave concern that must be satisfactorily and convincingly addressed before the manuscript is accepted for publication.

> *The Associate Editor made thoughtful suggestions for revising the notation in Section 4.2. We have adopted these suggestions as discussed below. We assure the Editor that formulas used in all computations is correct, and had already reflected the corrected notation in the previous submission.*

2 Associate Editor Comments

AE.0: For Section 4.2, my question starts with the definition of $\Gamma_{Z_{j^{(s+1)},j}}$. Is this quantity always defined for possible values of Z? That is, if $Z_j^{(s+1)} = n_A + j$, what is the meaning of $\Gamma_{n_A+j,j}$?

> *The Associate Editor is correct that $\Gamma_{n_A+j,j}$ is not defined under this notation. We had attempted to remove all instances in which this notation might be implied, but it seems we did not catch them all. We thank the AE for their thoughtful suggestions in this regard.*

AE.1: This notation first arises in the definition of $n_p(Z)$, which I think could easily be changed to something along the lines of $n_p(Z) = \sum_{j:Z_j \leq n_A} I(\gamma_{Z_j,j} = h_p)$.

> *Thank you for this suggestion. We have adopted this change.*

AE.2: My bigger concern comes for equation (18), where I am not entirely sure what probability is being computed, or how this sampling procedure combined with equation (19) would result in a "non-match" where $Z_j = n_A + j$. I believe that the idea is to partition the possible values of $Z_j : \big(\{r_{p_j}\}, n_A + j\big)$. Then, in equation (18), the left hand side is intended to be the probability that Z falls into each part of the partition. But, in the revision, I didn't see where the last "non-match" part of the partition is defined, and so there is no real way to select the non-match from equation (19).

> *The AE's understanding of the sampler is correct. The AE also correctly identified an error in equation (19), such that there was no real way to select the non-match option. We thank the AE for the recommended change in notation, and have adopted the recommendation. The sampling equations are now provided as follows:*
>
> *Define r as an arbitrary set of records. We have*
>
> $$p\left(Z_j^{(s+1)} \in r \mid \tilde{\gamma}, \boldsymbol{m}^{(s+1)}, \boldsymbol{u}^{(s+1)}, \pi^{(s+1)}\right) \propto \begin{cases} \frac{\pi^{(s+1)} N_{p_j}}{n_A} w_p^{(s+1)}, & r = r_{p_j}; \\ 1 - \pi^{(s+1)}, & r = \{n_A + j\}. \end{cases} \quad (1)$$
>
> *Since all records in A sharing the same agreement pattern with B_j are equally likely, we then sample among candidate records uniformly using*
>
> $$p\left(Z_j^{(s+1)} = q \mid Z_j^{(s+1)} \in r, \boldsymbol{m}^{(s+1)}, \boldsymbol{u}^{(s+1)}, \pi^{(s+1)}\right) = \begin{cases} \frac{1}{N_{p_j}}, & r = r_{p_j} \text{ and } q \in r; \\ 1, & r = \{n_A + j\} \text{ and} \\ & q = n_A + j. \end{cases} \quad (2)$$

AE.3: Also, it's not clear to me as currently written that the authors are equating the patterns with that partition so that by definition a "true" match that does not appear in A cannot have a pattern that appears in $\mathcal{P}$, but did not happen to be observed (because the match was not captured in A).

> *We did not quite understand this part of the comment. If record B_j has no match in A, then $Z_j = n_A + j$. If there is no matching record, there is no comparison vector (or pattern) for such a non-existent record.*
>
> *Despite the confusion about this specific sentence, the proposed change in notation more explicitly shows how the sampler handles non-matches, which we think is the AE's primary concern.*

AE.4: Page 9, just below equation (17): Should the reference to equation (9) should be updated to equation (13)?

> *The reference to equation (9) was meant as a comparison with equations (18) and (19), not equation (17). We have revised the paragraph to make this more clear. This material now reads:*
>
> *Note that sampling from (18) has complexity $O(P + 1)$ and sampling (uniformly) from (19) has complexity $O(1)$, regardless of the size of r_{p_j}. In contrast, sampling Z_j from the full conditional provided in (9) has complexity $O(n_A)$, because sampling a value from n_A options with unequal weights requires normalizing the weights to probabilities.*

AE.5: Supplement A, just below equation (3): Should the summation be over index j?

> *Yes. We have made the correction.*

AE.6: Supplement A, equation (12): I believe that this intermediate result is not described correctly. Because c_j is a function of γ_{ij}, once the function is divided by c_j, the quantity is no longer proportional to the probability. Rather, it is proportional to a likelihood-like function. Because equation (13) is conditional on γ, this distinction is not important for the end result.

> *We have removed the intermediate result in equation (12).*

References