Referee Report: "Privacy-Preserving and Lossless Distributed Estimation of High-Dimensional Generalized Additive Mixed Models" by Schalk, Bischl and Rügamer

1 Summary

This manuscript presents a distributed, privacy preserving component-wise gradient boosting algorithm for generalized additive mixed models. The authors propose a distributed row wise tensor product to account for site-specific effects. Their case study shows that the distributed algorithm provides equivalent model estimates compared to the model estimates based on the pooled data.

This paper presents an intriguing real-world application, yet several aspects necessitate attention. The key points for improvement are outlined below.

2 General Comments

1. Incorporating propensity score matching before applying the proposed algorithm could be valuable, especially if all data are derived from observational studies. This step might help alleviate unusual site effects (Hungarian)

2. To ensure coherence, maintain consistent notation for the penalty matrix Kl. In Equation (3), it's noteworthy that there's a smoothing parameter $\lambda_{l\times}$ preceding K_l, while K_l itself inherently contains a smoothing parameter.

3. Considering that the authors use the same number of basis functions, define $d_l = d$ for simplicity.

4. Some notations require clarification. For instance, the authors define $g_l(\mathbf{x}) = (B_{l,1}(x_j), \ldots, B_{l,d_l}(x_j))$ for spline base learners. Since $g_l(\mathbf{x})$ is contingent upon x_j,

including a subscript j in the notation is imperative. Additionally, for row-wise tensor product learners, the authors establish $g_j(\mathbf{x}) = (g_{j,1}(x_j), \ldots, g_{j,d_j}(x_j))$. To enhance clarity, it's recommended to consistently use $g_j(\mathbf{x})$ in lieu of $g_l(\mathbf{x})$ throughout the paper.

5. Due to the authors' consideration of interaction terms, it's worth addressing whether the proposed method adheres to the heredity constraint, which maintains main effects alongside their corresponding interaction terms within the model. See Wu and Hamada (2021).

6. Please report the computing run time for each method in Section 4.

3 Specific Comments

1. Page 2, Right side, line -6: (Li et al, 2020b) should be Li et al. (2020b). Use \citet

2. Define the abbreviation HP (Hyper Parameter?) upon its first usage.

3. Page 6, Right side, line 6: Clarify the objective of estimating random effects, as the primary interest typically revolves around estimating the variance components of these effects.

4. In Equation (3), consider relocating $K_{l\times}$ to the subsequent line.

5. Page 12, Left side, line -10: Specify the criterion employed for removing covariates. If correlation-based, elucidate the correlation threshold for covariate removal and the criterion for handling missing values.

6. In Fig. 1, how do the authors calculate the proportion of added base learners for each covariate? Should the proportion be the same at a given iteration number?

7. In Section 4, explicitly mention the covariates included in the final model as determined by the proposed method.

8. References: Gaye et al. (2014) International journal of epidemiology should be International Journal of Epidemiology.

9. References: McCullagh P, Nelder JA (2019) Generalized linear models. should be McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edition

10. References: There are two URLs for Rügamer et al. (2018) and Zhu et al. (2020). Please use just one.

11. Peferences: Add a journal name to Samarati and Sweeney (1998). k-anonymity: a model for protecting privacy. *Proceedings of the IEEE Symposium on Research in Security and Privacy (S&P). May 1998, Oakland, CA.*

References

- Wu, C.F.J and Hamada, M. S. (2021). *Experiments: Planning, Analysis, and Optimization, Third edition,* Wiley.