Review of *A general framework for unbiased mixed-variables distances*

Recommandation: Major revision.

While the manuscript contribution is not significant, it nonetheless addresses an important research gap in the literature, making it worthwhile. In its current form, the manuscript is difficult to read, somewhat incomplete, and the contributions are not presented clearly. If the following major comments are thoroughly addressed, the paper could be acceptable for publication in *Journal of Computational and Graphical Statistics*.

1. I believe the title of the paper is misleading. Calling it a general framework might oversell the scope, as the main contribution is a weighted mixed-variable distance. I recommend a shorter, more accurate title such as *An unbiased mixed-variable distance*.

2. An important portion of the paper, notably Section 3 and 4, primarily surveys existing distances that are compatible with the general distance. As a result, the manuscript reads partly like a survey synthesized around a new, more general distance. I suggest streamlining Sections 3 and 4, and expanding Section 2. In particular, consider formulating Definition 2.1 as a general function and then proving that it is a distance.

3. Regarding the previous comment, I am slightly confused by Definition 2.1. For two vectors $\boldsymbol{x}_1 = (x_{1,1}, x_{1,2})$ and $\boldsymbol{x}_2 = (x_{2,1}, x_{2,2})$, in the expression $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = d_1(\boldsymbol{x}_1, \boldsymbol{x}_2) + d_2(\boldsymbol{x}_1, \boldsymbol{x}_2)$, what exactly do $d_1(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $d_2(\boldsymbol{x}_1, \boldsymbol{x}_2)$ represent? Are they vector distances that emphasizes the j-th variable? If it is the case, could an example be provided, where the variable specific distances *take into account associations between the p variables*?

4. In this work [1], the authors proposed a mixed-variable distance in a more general hierarchical setting, *i.e.*, it can compare points that do not share the same variables. Its meta distance (Theorem 4) also uses scaling parameters, as in the present work. Given that the meta distance targets more general settings, what is the difference and advantages of the proposed distance? This paper should be discussed in the literature review, as it presents similar expressions and is already published.

5. The introduction is incomplete and would benefit from reorganization. On page 3 line 32, the motivation/objective is stated, however there are alternatives in the literature. What makes your work different and important? What are the applications motivating the work? The research gap should be stated explicitly, for example: *Although, many mixed-variable distances exist, they suffer from importance bias from variable types [...].* Also, the introduction contains a considerable amount of literature. For clarity, it is suggested to follow the context-problematic-research gap-objectives/contributions structure for the introduction, and moving the literature review content in a designated section.

6. The notation is difficult to read. Developing a clear and readable notation for mixed-variable problems is challenging. However, some equations are difficult to follow, *e.g.* Equation (2) contains two levels of subscripts. The notation should be modified to avoid such situations. I recommend using vector pairs $\boldsymbol{x}, \boldsymbol{y}$ or $\boldsymbol{u}, \boldsymbol{v}$ for definitions and expressions. Moreover, a designated notation section before Section 2 could be beneficial for stating the conventions used.

7. For the numerical experiments, the proposed distances 8) and 9) should be benchmark d on practical tasks, *e.g.*, regression or classification problems. Such experiments could better showcase the usefulness of the work.

Minor comments

The following minor comments, suggestions, and questions are intended to improve the clarity and flow of the manuscript, as well as to address minor errors.

1. P.2 line 14. What does it mean "*should not trivially impact the overall distance*"? This is not clear and inviting the reader to start the paper with confusing statements.
2. P.2 line 43. I understand that the term aggregating concerns combining variable specific distances, however for some readers it could be confusing with dimensionality reduction techniques. I suggest avoiding this term before it is introduced.
3. P.3 line 6. Is it too early to discuss code packages in the introduction? Would it be more appropriate to discuss them in the numerical experiments?
4. P.3 line 17. The statement "*the number of categories can affect the calculation of distances*" should be further explained.
5. P.5 line 47. Equation (2) is hard to read, although conceptually simple. Also, for the numerical variables, the subscript n might be confusing with dimension or number of variables for some readers.
6. P.7 line 38. How is it ensured in Definition 2.1?
7. P.7 line 52. "*The distance due to the numerical variables can be expressed as*" is not clear. Is it a specific instance of the proposed distance, or an equivalent statement?
8. P.8 line 8. Should it be $x_{i,j}$ on the right?
9. p. 10, line 43. What is the key takeaway from Section 3.1?
10. P.12 line 23. "*not uncommon*": double negation should be avoided.
11. P.17 line 27. For the reader, I recommend stating explicitly that methods 8) and 9) are from the work.
12. P.19 line 23. Could you explain why leaving one variable influences the distances of other categorical variables?
13. P.20 line 1. The desirable outcome is to have a horizontal line. This could be stated explicitly for the reader.
14. P.22 line 26. I recommend putting a horizontal line at the baseline average to show that this is the target, *i.e.*, the closer the better.

References

[1] Edward Hallé-Hannan et al. "A distance for mixed-variable and hierarchical domains with meta variables". In: *Neurocomputing* 653 (2025), p. 131208. DOI: 10.1016/j.neucom.2025.131208.