Referee report on Manuscript Number SIM-22-0663, entitled "Hypothesis tests for multiple responses regression: effect of probiotics on addiction and binge eating disorder"

The authors propose Wald tests for parameters of multivariate covariance generalized linear models. They assess the performance of these tests by means of computer simulations, and they analyze real data from a clinical trial that aims to evaluate the effect of the use of probiotics in the control of addiction and binge eating disorder in patients undergoing bariatric surgery.

Critical points:

- The methodological contributions of the manuscript are in my opinion rather limited:

  (i) Section 3.1 of the present manuscript, which is the main theoretical basis for the proposed tests, is essentially a literal repetition of Section 3 in the cited Reference No. 9 (Bonat and Jørgensen, JRSSC 2016).

  (ii) Utilizing a central limit theorem to propose/justify Wald tests is a standard approach in asymptotic statistics. See, for instance, Example 15.6 in the book by van der Vaart (1998, `https://doi.org/10.1017/CBO9780511802256`).

- At some occasions, the presentation lacks mathematical rigor. In particular, it is claimed on Page 7, lines 19-20, that the Wald statistic W follows a chi-square distribution with s degrees of freedom. There are several imprecisions involved in this statement:

  (i) The statement only holds under the null hypothesis.

  (ii) For non-Gaussian data, the statement only holds asymptotically as sample size tends to infinity.

  (iii) The statement only holds if the matrix $\mathbf{L}$ has rank s.

- The simulation study is not convincing (to me). In particular, in Figures 2 and 3 the region around the zero distance (leftmost part of the figures) should be zoomed, so that one can see how accurately the test keeps the significance level for finite sample sizes. Namely, several authors have noted that the Wald test is prone to behave rather liberally for finite sample sizes; see, among many others, Konietschke et al. (2015, `https://doi.org/10.1016/j.jmva.2015.05.001`), and the references therein. A bootstrap-based calibration of the Wald test has

been recommended to remedy this issue. I am quite sure that the same issue arises in the authors' (MANOVA) context here.

- I disagree with the statement that it is usual in tests of multiple comparisons to recommend the Bonferroni correction (Page 11, lines 52-53). What exactly do the authors mean by "usual" here?

- Language and grammar in the manuscript need substantial polishing. Let me give only two specific examples:

  (i) In the very first sentence of the summary, it is unclear whether "where" refers to "clinical trials" or to "medical research".

  (ii) The sentence "However, for multiple response regression models, there are few alternatives for hypothesis testing." (Page 3, line 4) is hard to understand (for me). Do the authors mean that the choice of the alternative hypothesis is limited, or do they mean that only a handful of tests exist?

  In general, I recommend to consult a native speaker for help with language and grammar.