Review report for “Inside Out: Externalizing Assumptions in Data Analysis as Validation Checks”

This paper presents a procedure for the formal analysis of validation checks in data science applications, which has the potential to help data scientists identify potential issues more systematically. This is a valuable and timely effort that could benefit the broader data science community. The paper is generally well-written and easy to follow. It could make a meaningful contribution to the field, provided that the comments below are adequately addressed.

General Comments

1. The manuscript is being considered for the special issue on Statistical Aspects of Trustworthy Machine Learning. However, the connection to the theme of the special issue feels somewhat underdeveloped. It would be helpful for the authors to clarify which aspect(s) of trustworthiness are being addressed. While the work appears to relate to robustness and reproducibility, these connections should be more clearly articulated and substantiated in the manuscript.

2. The proposed procedure appears to assume that it is possible to simulate datasets similar to the observed data (e.g., as illustrated in Figure 2). However, this assumption may not be realistic in many practical scenarios. Simulating data requires a data-generating model, but commonly used statistical methods such as generalized estimating equations (GEE) or the generalized method of moments (GMM) do not fully specify the underlying distribution or likelihood. Even when the mean structure is specified, there may be infinitely many likelihood models consistent with it. Simulating from only one such model risks overlooking a broad class of plausible alternatives. How do the authors propose to handle such cases? More broadly, generating synthetic data with properties similar to real data is a fundamental and challenging problem in data science practice. The authors may wish to comment on recent advances in this area that aim to address this challenge without fully specifying a probabilistic model.

3. The construction of validation checks appears to be a critical step in applying the proposed methodology. It would be valuable if the authors could offer some guidance or recommendations on the principles for designing these checks. Clarifying this aspect would enhance the practical utility of the approach and help readers apply it more effectively in real-world settings.

4. The independence metric appears to be ill-defined or uninformative for single checks, much like defining total correlation for a single variable. By definition, the metric seems to always take the value one in such cases, as observed in Tables 1 and 3. Including it in the total score may artificially inflate the overall score when compared to checks involving multiple components. Further, in the current examples, the multivariate checks consistently yield independence values close to one. Are there cases where this metric meaningfully deviates from one? Clarifying the interpretation and role of this metric, particularly in the single-check case, would strengthen the methodology.

5. The writing is generally clear and easy to follow. However, there are a number of minor syntax errors that could be addressed with more careful proofreading. The use of AI-based writing assistance tools could be helpful in this regard and, in my view, would constitute an ethical and constructive application of such tools.

Specific Comments

1. Section 4, paragraph 2: Should the "simulated datasets" be independent of the observed data? If so, this is a big difference from bootstrap samples and a remark would be useful.

2. Figure 2: Quadratic mean is mentioned in the text but not shown in the SCORE box.

3. Section 4.1, paragraph 1: $X_1, \ldots, X_n$'s are undefined at the first occurrence; n means the number of checks, but in statistics it is often used for sample size.

4. Section 4.2, end of paragraph 1: I assume the dash should be the minus sign. Having a minus sign in a sentence is confusing. This needs to be rephrased, perhaps just use "minus".

5. Section 4.2, paragraph 2: What is the range of total correlation? My understanding is that it is $[0, \infty)$. Even after standardizing to per observation, the range may still be not $[0, 1)$ as desired.

6. Continue: p is undefined; the equation needs to close with a period.

7. Table 2: How did the authors come up with these checks? Readers would appreciate the motivations of these checks. Some of the checks are apparently correlated (e.g., the four checks on mortality-temperature correlation).

8. Section 5.1, paragraph 1: Gaussian copula does not allow tail dependence. So by imposing a Gaussian copula, it is assumed that there is no tail dependence in these three variables. As questioned in one of my general comments, this might not be desirable.

9. Figure 4: Could add bands of 95% confidence intervals of the QQ-plot; the texts on top of each panel could be moved to the caption; the labels of the axis should clarify empirical versus theoretical.

10. Figure 5 caption: The negative sign in front of 0.03 needs to be fixed.

11. Section 5, paragraph 1: The “our second example” appears to be a continuation of the same analysis.
12. Table 4: There are four checks for the mortality-temperature correlation, but only one check reported in this table.
13. Section 5: It’d be valuable to see some real improvements from the checks. For example, fit a poor model first and identify the issues from the checks; then fit a strong model to fix the issues and pass the checks.

Minor Comments

1. Abstract: “simulations of the original data” is misleading.
2. Page 2, paragraph 1: “thought” process, not “throught” process.
3. Page 2, paragraph 2: “is measures”, two verbs.