Response to Reviewers JDS24112

Inside Out: Externalizing Assumptions in Data Analysis as Validation Checks

H. Sherry Zhang, Roger D. Peng

2025-07-08

We would like to thank the reviewers and editor for taking the time to review our paper and provide constructive comments. Below is a point-by-point explanation of how we have changed the paper. Reviewers' comments are in normal text and **our responses are in bold text**.

Reviewer 1

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. **A: We now discuss the connection between our work and the trustworthiness of machine learning in the third paragraph of the *Introduction*. Specifically, our work address the breakdown in trust caused by violations of algorithmic assumptions – corresponding to the third step in Broderick et al. (2023). In modern data analysis, where data are often not collected directly by the analysts, assumptions about the data can be unknowingly violated, leading to unexpected outcome. Our approach advocates for making these assumptions – often simple or taken for granted – explicit, in order to more efficiently diagnose unexpected outcomes.**
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. **A: The reviewer is correct that in certain scenarios (e.g. GEE and GMM), it can be challenging to accurately simulate dataset that resemble the observed data. However, the primary aim of our paper is not to recover the true data generating process or to find the model that best fits the data. Instead, our goal is to understand which characteristic of the data may lead to the unexpected results - a mismatch between the data and the model. Here, we hold the model fixed and examine the idiosyncrasies in the data that could causes such incompatibility. This contrasts with the approach of treating data as fixed and focus on improving the model. Our perspective is motivated by the applied studies, where certain modelling strategies are generally accepted for specific types of data. When unexpected outcomes arise, analysts often first investigate whether particular characteristics of the data may cause the results, rather than developing new models. In this context, the analysis validation checks we propose provide a pre-developed "diagnostic checks" to investigate the data. To note that, even when the results is as expected, we don't necessarily assume the model is correct. For generating synthetic data with similar properties to real data, we discussed in the discussion the potential of using recent work in data thinning, data fission and differential privacy.**

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. **A: In the toy example section (Section 3.1), we include some explanations on the construction of validation checks: these are diagnostics analysts may naturally check when they receive an unexpected outcome. For example, when the mean of a vector falls outside an expected range, analysts may check the the five-number summary or plot the variable distribution. The validation checks are those diagnostics framed as binary predicates.**

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. **A: After reviewing the definition of the independence metric, we find it is the previous version is less ideal and we have refined the definition. The amount of overlapping information among checks can be represented by the normalized total correlation,** $C' = \frac{C(X_1,\cdots,X_k)}{\sum_{i=1}^{k} H(X_i)} = \frac{\sum_{i=1}^{k} H(X_i) - H(X_1,X_2,\cdots,X_k)}{\sum_{i=1}^{k} H(X_i)}$**, where**$H(X_i)$ **is the entropy of check** X_i**,** $C(X_1,\cdots,X_k)$ **and** $H(X_1,\cdots,X_k)$ **are the total correlation and joint entropy of checks** $\{X_1,\cdots,X_k\}$**, respectively. The amount of independent information is then** $\eta = 1 - C' = \frac{H(X_1,\cdots,X_k)}{\sum_{i=1}^{k}}$**. The independence metric is now defined as** $\frac{\eta - 1/k}{1-1/k}$ **to scale** η **to [0, 1], where 0 is attained when all checks are independent and 1 is attained when checks are all identical. A trivia case occurs when there is only one single check, then both the joint entropy and the sum of all entropies is the entropy of the check itself, hence giving an independence of 1. These changes have been reflected in Section 4.2 and in the examples.**

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. **A: We have run though the spelling check on the article and fixed the spelling mistakes.**

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. **A: The simulated dataset in our paper are not necessarily independent of the observed data, and we do not rely on bootstrap sampling at any point, although it can be a method to generate the simulated data. We have include a discussion of the principle and different strategies to generate the simulated data after paragraph 2 in Section 4.**

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

3. Section 4.1, paragraph 1: X1, . . . , Xn's are undefined at the first occurrence; n means the number of checks, but in statistics it is often used for sample size. **A: We have added the definition of** X_i **as the binary predictors in Section 4.1 and replace** n **with** k **for the number of checks.**

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". **A: We have replaced the dash with the word "minus" in the text.**

5. Section 4.2, paragraph 2: What is the range of total correlation? My understanding is that it is [0, ∞). Even after standardizing to per observation, the range may still be not [0, 1) as desired. **A: We have re-defined the independence metric (See *General comments* #4) and the metric is now bounded in [0, 1].**

6. Continue: p is undefined; the equation needs to close with a period. **A: We have added the definition of p(.) and close the sentence.**

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). **A: Four mortality-pm10/ mortality-temperature tests are constructed because analysts may not be sure what the best cutoff point for the correlation tests would be. This example illustrates how the method developed can be used to determine the cutoff value for tests.**

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. **A: While using a different copula (t- copula or extreme value distribution copula) to better model the tail dependence could produce better simulated data to model temperature and mortality, high temperature does not necessarily correlates with high PM10 in air pollution literature. We have maintained the current simulation scheme but add a discussion on the matter in the discussion section.**

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. **A: We have added the 95% confidence interval, remove the plot title to the caption, and refine the axis labels.**

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

11. Section 5, paragraph 1: The "our second example" appears to be a continuation of the same analysis. **A: We use the wording "second example" because we count the toy example as the first example. But, to avoid this confusion, we have changed the wording to "In the following example".**

12. Table 4: There are four checks for the mortality-temperature correlation, but only one check reported in this table. **A: addressed in *Specific comments* #7.**

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. **A: The reviewer's comment assumes we need a better model when an unexpected outcome occurs, however, in practice, we always assume there is something issue with the data, rather than rushing into developing new methods. The analysis validation checks are used to motivate a better understanding of the data, not an ideal model (This point is also addressed in *General comments* #2).**

Minor comments

- Abstract: "simulations of the original data" is misleading. **A: we remove this information in this sentence for easier understanding of the sentence.**
- Page 2, paragraph 1: "thought" process, not "throught" process. **A: Fixed.**
- Page 2, paragraph 2: "is measures", two verbs. **A: The redundant "is" has been removed**.

Reference

Broderick, Tamara, Andrew Gelman, Rachael Meager, Anna L. Smith, and Tian Zheng. 2023. "Toward a Taxonomy of Trust for Probabilistic Machine Learning." *Science Advances* 9 (7). https://doi.org/10.1126/sciadv.abn3999.