Reply to reviewers

We thank the Associate editor and anonymous reviewers for the feedback on the paper. In discussion with the editor, it is agreed that becuase the article was submitted as a *short technical note* that the revision would focus on these changes only:

1. Make all the small corrections listed by Reviewer 2.
2. Check the computational details to ensure that it is clear.
3. Check the literature references, to ensure nothing major is missed. The reviewers didn't suggest articles that we had missed citing. We can try to make the context clearer.
4. Ignore the request for a broader comparison or add more examples. The two examples provided are contemporary and differ in nature, illustrating general usage.

Reviewer 1

1. Providing a more comprehensive review of existing methods would enhance the context for the proposed index.

The introduction was slightly expanded, and two new references added to provide additional context.

2. A detailed comparison with existing methods, emphasizing how the new index offers improvements, would clarify the study's rationale. Including a comparative analysis with current projection pursuit methods, while highlighting the strengths and potential limitations of the proposed index, would further strengthen the paper.

This is first work that allows the comparison of a reference normal distribution to observations using projection pursuit, and a projection pursuit guided tour. New text in the introduction helps clarify this.

3. Including additional real-world data examples would enhance the practical relevance and robustness of the findings.

The two examples used in the paper illustrate different uses of this work in practical settings. The first example uses simulated data for privacy purposes but the data reflects patterns seen in clinical data. The text has been modified to make this clear. There is no room in a short technical report to add more examples.

4. The methodology is well-explained, but providing more details on the computational aspects and potential limitations would improve reproducibility. Clarification on how to obtain and interpret the "index" results is needed. Offering more detailed descriptions of the algorithms and their implementation would aid in understanding. A deeper discussion of the practical implications of these results would add valuable context.

We have extended the implementation section to mention the options for optimization of the index, and have included a reference for the ferrn package that can be used to investigate index values and to compare different optimization strategies. Including such an investigation would be beyond the scope of a short technical note. All implementations are available in CRAN packages and the code to reproduce the examples are available on GitHub for full reproducibility.

Reviewer 2

The paper's ideas are interesting, but some sections of the article are very brief and summarized, making it difficult to understand the method. Additional references to similar works are needed to better justify the main contribution of the work. It is also necessary to number the main equations and reference them in the text when appropriate. More details in the pdf.

The introduction has been expanded slightly with two more references added to provide more context.

We have added numbering to the main equations such that they can be referenced in the explanations, this should make the derivation easier to follow. More detail has been added to the theorem and proof in Section 3.

The proposed method addresses a similar question to that of a normality test for a specific data sample, but it can also be used as a method to detect outliers—data that were generated from a distribution other than the reference multivariate normal. In this regard, what differentiates your method from similar approaches, and why do you propose it as an improvement over existing methods? I believe it would be helpful to include additional context and references on other projection pursuit contributions aimed at finding interesting projections in the same direction as your approach.

We have additional context and references in the introduction.

In page 2 line 47: "This paper describes a new projection pursuit index which is optimized by projections where a new sample is most distant from the existing normal distribution." At this point, it would be beneficial to emphasize the significance of your findings and highlight the main contribution of your paper. Clarifying how your method advances the field or outperforms existing techniques will strengthen your argument and provide a clearer understanding of the value of your work.

We have added a short review of traditional outlier detection methods and relevant methods here, and emphasized where they currently do not apply.

Page 3 line 26: Could you explicitly define the shape transformation mentioned in point 3?

We have rephrased points 2 and 3 to make this more clear.

Page 3 line 31: Are the new observations transformed in any way before being plotted for comparison? If so, could you clarify what kind of transformations are applied and how they impact the comparison with the reference distribution?

The new data needs to be pre-processed in the same manner as the reference distribution, we have added this information in the description. For example, if the reference distribution is scaled using means and standard deviations of historical data, these same means and standard deviations need to be used to scale the new data.

Page 3 line 57: "Figure 2 compares a new sample of patient scores against the normal range", However, there is no prior context provided for this example, making it unclear to the reader. Additionally, on page 4, you mention "normal patients" as if the reader is already familiar with the patient example. It would be helpful to introduce and explain the patient data context earlier in the paper to avoid confusion.

The text in section 2 has been changed to remove any specialist language. The explanation is also more detailed now.

Figure 2: The labels in the figures are not clear. In panel a), the label "ci" is confusing, and in panel b), the label "norm" may lead to confusion regarding normal patients. To enhance clarity, consider increasing the point size and applying transparency, similar to the size and transparency used in Figure 1.

This figure has been modified to make it clearer, and to correspond with the changed text in the section.

Section 3: The flow of the content is difficult to follow, and some equations would benefit from being described in a different order. Additionally, numbering some of the equations and referencing them in the text would significantly improve readability.

We have numbered the main equations and now use referencing in the description. Segments of this section have been rewritten to include more explanation and enhance flow, and variables have been defined earlier, before they are used.

Page 4 line 57: Should include "Let x a p dimensional vector"

This has been added.

Page 5 line 7: Please number the equations and reference them in the text (for example in line 16 and 43). This will enhance clarity and make it easier for readers to follow your arguments

Equation numbers and references have been added.

Page 5 line 16: "Theorem. The projection of this p = D ellipsoid in 2-D has the equation" you should change "this" for "a p-D ellipsoid from equation 1" or something similar. Additionally, after the theorem, please explain the variables mu, p, P and y. While you provide some of this information in line 41 and on page 6, it would benefit from better organization for clarity. Furthermore, consider restructuring the proof to enhance its clarity.

The ellipsoid equation is now specifically referenced in the statement of the theorem. The mentioned variables are now defined before the statement of the theorem, helping reorganise the section to improve clarity. Although the proof is not fundamentally restructured, it has been reworded with further explanation added.

Subsection 3.2 is unclear and would benefit from additional details: Page 6 line 34: "To define a measure of an interesting projection is to maximize the average Mahalanobis distance", you should rephrase "is to maximize". Additionally, since the Mahalanobis distance serves as a fundamental tool for detecting multivariate outliers, it would be beneficial to mention this and provide a definition of the distance before explaining your method. Moreover, I recommend considering a robust version of the Mahalanobis distance, as this could enhance the robustness of your analysis.

We have rephrased the sentence and included a definition of Mahalanobis distance. Robust methods are used when estimating a reference distribution from a data sample containing outliers, thus implicitly we are using a robust version of the distance.

Page 6 line 55: Please number the equation and explicitly state that this is the new projection pursuit index you are proposing.

Done.

Page 6 line 37: You mentioned W but it would be beneficial to define it here. While you provide this information on page 7, including a definition in this context will enhance clarity for the reader.

We have moved the default definition for W to where it is first mentioned.

Page 7 subection 3.3 It would be useful to include a simulation example demonstrating how to apply the method in cases where the observations deviate from the norm in different directions. This example should illustrate how to group these observations effectively, providing readers with a clearer understanding of the practical application of your method. If you address this point later in the paper, please mention it here to guide the reader.

Including a detailed simulation study is beyond the scope of a short technical note. However, the second example shows how clustering can be used in an application. This is now mentioned in Section 3.3.

Page 8 Figure 3: You can justify that your method is useful for gaining insights into the outliers in the data. Emphasizing this point will help underscore the practical significance of your approach and its role in identifying and understanding anomalies within the dataset.

We have emphasized that point in the text referring to Figure 3.

Page 9 line 18: "There are normal ranges" maybe acceptable ranges is better

Done

You should move Figure 4 to page 10. In the figure caption, please correct the phrase "Red cross indicates observation is outide the 4-D confidence ellipse" to fix the spelling of "outside" and 2-D.

We have moved Figure 4 up (now page 12), and fixed the spelling.

Page 12: All your comments regarding the clusters are unclear because Figure 5 lacks color labels. Including these labels will help clarify your points and enhance the reader's understanding of the clustering in the visual representation.

A color legend has been added in Figure 5a.

Page 13: In Figure 5, the selected color palette is not effective, making it difficult to distinguish the five clusters. Additionally, using white for one of the groups against a white background further complicates visibility. I recommend trying the Dark2 palette or another contrasting palette. Please also include color labels to enhance clarity and help the reader identify the different clusters easily.

We have updated the color palette and now use shape to de-emphasize points that are not outlying (cluster 0). The color legend has been added to the Figure.

Page 14 Figure 6, In Figure 6, please change the cluster number colors to match those in Figure 5. The cluster numbers are difficult to see, so consider increasing the font size or other option. Additionally, you should include a color legend for clarity. When describing the variables in the text, please include their code names in parentheses. Also, ensure that the variable names in Figures 5 and 6 are consistent for better coherence throughout the paper.

The colors are matched between Figure 5 and 6. We have increased the size of the cluster labels in Figure 6 to make them more readable, and have added a color legend. Variable nams in Figure 6 have been matched to those in Figure 5, and are now always added in parentheses when mentioned in the text.

Page 14 line 48: Is Shapley Cell Detector algorithm defined in Mayrofer and Filzmoser(2023)?

It is, we have added the reference also to where the algorithm is mentioned to make this more clear.

Page 15 line 19: In the conclusion, you mention that your work relates to outlier detection methods, particularly those using robust statistics. However, this important point is not clearly explained in your paper until the examples. It is essential to clarify this relationship earlier in the paper, before the examples, to provide context for the reader. Additionally, in the conclusion, you include some references that could be incorporated earlier in the text. This would help justify the differences between your method and existing ones, enhancing the overall clarity and depth of your discussion.

We have extended the introduction somewhat to include references to outlier detection in the context of linear dimension reduction. The additional references in the conclusions show potential generalizations of the work connecting to other directionsin outlier detection.