Review of Stamatis et al: “Development and Application of a Supervised Pattern Recognition Algorithm for Identification of Fuel-Specific Emissions Profiles”

General comment

The manuscript introduces interesting algorithm for identification of different fuel types from emission profiles. The algorithm is a combination of widely applied statistical methods which together make a good tool for emission classification. However, the methods need to be described with more detail and the manuscript needs to be structured better. The methods should be separated from the results, now the text is difficult to read as new methods pop out from nowhere in the middle of results. With these and the detailed comments below properly addressed I could recommend the manuscript for publication.

Detailed comments:

Page 4 lines 115-118: IS the reason for missingness always the compounds being below the detection limit? If there are missingness due to e.g. instrument malfunction, filling in the zeros might bias the further analysis.

Page 4, lines 119-122: Feature selection needs a bit more clarification. ANOVA is sensitive to non-normality and heteroscedasticity of the data and if this is not taken account the feature selection may be biased. Whereas for PCA, standardization should not be standard procedure (Gewers et al. 2018), especially if the components are calculated with singular-value decomposition (SVD, Isokääntä et al. 2020) as I assume was applied here since the authors refer to Abdi&Williams 2010 paper. In addition, it should be stated in the methods section if SVD or eigenvalue decomposition was used to find the principal components.

The whole section introducing the PR algorithm should be extended to give more details on the methods used. It should be stated why PCA was chosen over to other dimension reduction methods, like explorative factor analysis, and similarly for LDA. PCA execution should be described with more details. In addition to decomposition method, it should be stated if some rotation method was used and which type of rotation: orthogonal or oblique and which version of them. Rotation methods are discussed in Abdi&Williams and in Isokääntä et al. (2020).

Section 3.2.1. The feature selection needs more justification. As PCA is, by definition, dimension reduction method, why the dimensions need to be reduced before running PCA for PR? Was the explained variance used as selection criteria for lower number of compounds in the analysis? It should not be as higher number of variables means higher total variance, which in turn leads to lower explanation rate with same number of components. Is the result in Fig 2. for the selected 5 (or 10) variables? I yes, then it should be noted that with five components you will explain 100% of the variation of five variables (but not for 10 variables). Since PCA tries to explain most of the variance with the first component, it is expected that the explanation rate is high. Naturally, with statistical models the aim often is to get as simple model as possible but the selection criteria need to be well reasoned.

Line 189. Fig 2 does not show normalized eigenvalues. Add the scree plot.

Line 195: Why Euclidean distance? I am not questioning the choice but wish to see the reasoning

Section 3.2.3: Did I understand correctly that you were only using one pair of components at the time in cluster analysis? If yes, I would ask why? Cluster analysis is a multivariable method, and thus can be applied even for all components at the same time. Figure 8 shows that PC4 does not really affect the clustering, but it is defined by PC1. Thus, this paired comparison seems redundant.

Page 13, line 243-244: Separation would be improved also by increasing the number of compounds in the analysis or using different rotation in PCA.

Results and Conclusions: Too strict limitations in feature selection also reflects to results and conclusions. Using more features in the model would probably lead better separation of sources in mixed samples but in creasing the number too much would decrease generalizability of the model or even lead to overfitting. Thus, it is important to find the balance.

References:

Gewers FL, Ferreira GR, de Arruda HF, Silva FN, Comin CH, Amancio DR, Costa LD. Principal component analysis: A natural approach to data exploration. ACM Computing Surveys, Volume 54, Issue 4, Article No.: 70, pp 1–34, https://doi.org/10.1145/3447755

Isokääntä, S., Kari, E., Buchholz, A., Hao, L., Schobesberger, S., Virtanen, A., and Mikkonen, S. (2020)
Comparison of dimension reduction techniques in the analysis of mass spectrometry data
Atmos. Meas. Tech., 13, 2995–3022, doi:10.5194/amt-13-2995-2020.