Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-469-RC2, 2019

Atmospheric Chemistry and Physics
Discussions

Interactive comment on "Dimensionality-reduction techniques for complex mass spectrometric datasets: application to laboratory atmospheric organic oxidation experiments" *by* Abigail R. Koss et al.

Anonymous Referee #2

Received and published: 28 September 2019

The authors compare three different techniques for dimensionality reduction in mass spectrometry time series data sets, positive matrix factorization, hierarchical clustering, and gamma kinetics parameterization. They evaluate the behaviour of the three techniques on two data sets and conclude that PMF is not competitive compared to the two others.

Overall the paper gives a good overview of the work, but requires some revision before being published.

For the clustering, the authors chose agglomerative clustering using average linkage. However, they don't provide all necessary information to reproduce the experiments or motivate their choices. Euclidean distance was used, why was this distance measure chosen? I can see it makes sense in some ways, however there are different metrics specific for time series, most notably dynamic time warping (DTW), which might make sense in this case. DTW compensates for shifts in the time series, so for particular use cases, this could make sense. Otherwise, there are more approaches to achieve a clustering, why chose this one? Agglomerative tends to be computationally faster than divisive, but this shouldn't be a problem with data sets this size. Otherwise what about density-based clustering or really simple approaches such as K-Means?

Similarly, there exist a wide range of algorithms for PMF, which one was actually used here? And why this one? The authors give the library, but some details on the method would be necessary.

On page 11, the authors give a formula for the quality of fit parameter Q, but half of the variables in the formula are not defined anywhere so the formula does not really make sense.

Also, when looking at algorithms such as PMF or clustering, it would be interesting to calculate performance measures and give them to get a feeling how well the clustering or factorization works. This could be simple reconstruction error or normalized mutual information (if there is a ground truth).

Another issue is repetition of experiments. While the agglomerative clustering should be mostly stable, PMF usually is not when using big enough data sets. So a single run would not be a reliable representation and multiple runs would be necessary. Additionally, this paper seems to base its results on two data sets, which cannot give any reliable or statistically sound performance representation for these approaches. Anything below at least 5 data sets won't give you the proof you need for what you state in the conclusion. Either rerun the experiments a lot more times or restate in the conclusion that this gives an indication, but to proof it, many more experiments would be needed.

I would recommend improving the presentation, particularly the figures. It is not always straight forward to understand what is shown. For example Figure 14 on its own does not explain the meaning of the colours or the size of the dots. Also the labeling of part A, B, and C is not very standard and sometimes hard to understand.

One specific question I have on Figure 2. I don't see how the cluster of C happens, looking at the data in C, this does not seem to be a cluster, the highest gray line is far away from all others and looks far too much as an outlier compared to the rest of the cluster. How does this compare to the rest of the data? Is everything else just even further away?

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-469, 2019.