Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-469-RC1, 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 #1

Received and published: 23 July 2019

Koss et al. present three statistical/mathematical approaches to reducing the time series of multiple species observed or simulated in a laboratory chamber into chemically meaningful groups or clusters. The authors conclude that the PMF (positive matrix factorization) technique does not perform nearly as well as the HCA (hierarchical clustering) or GKP (gamma kinetics) techniques in binning compounds (observed or simulated) into proper generational groups that share common chemical ages. The manuscript is concise, well written. Figures are illustrative, support the reported conclusions. This review points to a few areas of ambiguity that should be addressed/clarified.

The work is appropriate for publication in ACP after these minor revisions.

PMF does a poor job compared to HCA at assigning members into chemically meaningful groups. This result is a big deal given how widely PMF is used. More discussion and/or tests are needed to explain exactly why the PMF does not perform as well, that is, can PMF be modified so that it performs as well as HCA? Mathematically, matrix factorization and hierarchical clustering are similar. They certainly have the same intended goals. One difference is that with HCA each time series is normalized such that all compounds are more or less given equal weight, whereas PMF is biased towards those with higher signal to noise. Though it is not standard operation, it would be useful to re-run PMF on the ~400 TMB products but after normalizing each member such that they are given equal weight. Basically, feed the same input to HCA and to PMF. Let's compare apples to apples. Doing so will rule out differing inputs as the reason for the different performance. Also, constrain PMF such that each member can belong to only one factor.

It would be helpful to make mirroring figures so that comparing the performance of each of the three techniques is easier. For instance, make figure 7 for PMF look like figure 9 for HCA. Include a figure like figure 6b (mass spectrum of each factor) but for HCA results as part of figure 8. Same with GKP. Is the reason that PMF factors do not accurately represent chemical age groups is because each factor contains compounds with a wide range of amu (as shown in figure 6)? Is this not the case for HCA (please show in figure 8). And also for GKP.

The way that HCA is described on page 12, it reads as if the technique also solves for the final number of clusters needed to explain the variance of all input members. But it turn out (page 23 line 465) that this final cluster number is "chosen" by the user. How was this final number objectively determined? There are eight clusters (figure 8), but that can easily be reduced further (6, 7 and 8 look pretty similar, as do 3 and 4; and conversely, each of the those eight can be split even further). PMF at most has 6 factors. GKP has 9. How are the final group/factor/cluster number chosen for PMF and

GKP? Perhaps the authors should choose the same number for the three techniques. This again, I think will present a fairer comparison of the three techniques.

There are sections dedicated to PMF (section 3.1) conducted on both simulated and chamber data. Same with GKP (section 3.3). Why not HCA (section 3.2)?

Minor Figure 1 bottom. legends needed.

page 7, line 167: "Teflon" itself is PTFE and is trademarked and manufactured by a company called Chemours. Is the tubing PFA or PTFE? Manufactured by Chemours? If not, it is not Teflon.

Figure 2. Panel D. Not a great figure to highlight in main manuscript. Account for oscillation before including as main figure. I understand it is not included, but most people only look at figure and not read caption, will come away with wrong impression. Panel B not informative. This figure perhaps is introduced prematurely since hierarchical clustering since that section is far below.

Page 12 line 290, Need hyphen in citation

Figure 8, 9 and 14 share the same color scheme. It would be nice to have a common legend and/or color-bar shown in each of these figures to remind the reader that these colors represent generation determined by HCA.

Please make clear in each of the figure caption in the SI and main manuscript whether it is presenting simulated data or measured data.

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