Review of S. Isokääntä et al. (AMT-2019-404):

Comparison of dimension reduction techniques in the analysis of mass spectrometry data

Summary

Sini Isokääntä and co-authors present a comparison of dimensionality-reductive techniques for mass spectral data analysis, applied to a data set of car exhaust + a-pinene aging experiment in a reaction chamber. The data, collected by a PTR-MS and an AMS, was factorized and clustered using 5 different techniques. The authors present and discuss their analysis procedures and interpretation of results from mathematical and physicochemical viewpoints. For the PTR-MS data, all five techniques produce comparable results, yielding 4 to 5 interpretable factors (or clusters). This is a very important result considering the novelty of applying these methods to PTR-MS and similar mass spectral data of gas phase. For AMS data, only PMF and NMF yielded results that could be compared – PCA, EFA and PAM struggled with the particle data set. Although applied to data from a single experiment, the comparisons presented and the discussion on analysis techniques convey general messages for many similar analyses that abound in atmospheric science.

The manuscript is clearly of interest to AMT readers, and touches the important topic of computational analysis of complex, large mass spectrometric data sets and advanced statistical methods of data analysis. Despite the evident need, a very limited amount of reviews of advanced statistical methods for this type of complex, physicochemical analyses is available, so I see the authors' contribution as extremely welcome to the field. The manuscript is well and clearly written and structured. With some exceptions, detailed in specific comments, the data, analysis techniques and experimental setup is adequately described. I recommend the paper for publication in AMT after addressing the following comments.

Major comments

The authors generally do a good job of introducing and describing their methods, but I would call for discussion on the physicochemical *objectives* or *purpose* of the statistical techniques used in this study. Obviously it is about more than reducing the size or dimensions of the data, which is the single common feature of all these SDRTs. However, some of these methods are rather similar (e.g. PMF vs NNMF) while some aren't (e.g. PAM vs factorization methods). Generally, a data analyst should choose a proper tool for the job, depending on the goal. Here it seems the methods are just applied on the data, seemingly without stating beforehand e.g. the difference in categorizing or classifying variables using PAM, studying the main explanative components of variability or doing latent feature extraction by exploratory factorization. Many of these differences are later casually mentioned, but I suggest an effort to further summarize these fundamental differences, and conversely, some of the close similarities could be made in the introduction and conclusions of the paper.

Secondly: algorithmic methods are often discussed plenty in these types of reviews, but some of the equally significant, but less obvious issues include, among others: data pre-processing (e.g. scaling and weighting), metrics (e.g. for quality evaluation and similarity), and error models. Mostly, the authors cover these topics well enough, but some major questions surrounding e.g. the non-standard PMF error model the authors used and preprocessing of data for PCA/EFA are raised. See specific comments on the details.

Finally, most factorization methods, including those used in this paper, assume constant factor profiles, i.e. that chemical reactions do not take place. This is fundamentally at odds with the approach of using them to model data of reaction chamber chemistry. In practice, the methods *can work* despite this (as is shown in this work), but this fundamental violation of basic assumptions should be explicitly stated and discussed.

Specific comments

p.2.l.2.: "…sharp change at the beginning of experiments, e.g., switching on UV-lights) may present additional challenges, as PMF was originally developed for field measurement data sets where real changes in factors are expected to be much slower than e.g. the noise in the data.". In my understanding PMF is not assuming anything about order of measurement points or rates of change in loadings – please provide a reference or expanded rationale for this line of thinking!

p.2.l.14. "In our study we chose a set of SDRTs having fundamental differences." Aside from data reduction or simplification, the objective of SDRTs differ. Often, clustering is applied to classify or categorize non-correlated variables, whereas factor analysis aims to uncover latent features the combination of which would best explain the variation in data. Can the authors please comment on their *objective of the analysis* outside of reduction of data size) of the use of SDRTs for this type of analysis? Especially regarding clustering.

p.2.l.16: "On the other hand, clustering might be more suitable for a more simplified or preliminary approach, or when the chemical compounds in the data are already known." It is not trivial to envision a case where clustering would be preferable, even as a preliminary approach – please expand on this reasoning or provide an example.

p.3. : Please describe the PTR-MS and the SP-AMS in some more detail here. Specifically their mass analyzer resolution (C/HR/L –ToF), as this strongly affects mass analysis and type of data.

p.3.l.9.: A note. As the authors state, if one does not account for mass spectra baseline, this may give rise to "background factors" or cause variables to get polluted or mixed (depending on instrument resolution). While theoretically these could be separated by factor analysis, in practice their presence in data hinders the analysis, often notably. For hard clustering, this effect of may even more problematic. Can you give an order of magnitude estimate of this effect? Can you separate baseline factors / clusters in some cases? What is their fraction of total explained variability?

p.3.l.18.: This type of truncation correction is likely detrimental to analysis and should be avoided in statistical analysis (especially factor analysis, as it creates an artificial, non-random, variability component). If only few points had this problem, would it not be preferable to omit them?

p.5.l.30: "[…]*where X includes the analysed variables (centred and scaled by their standard deviations)*" In their examination of errors in PCA (and PMF) in environmental applications Paatero and Hopke (2003) strongly advice against "autoscaling":

> *"5.1. Do not autoscale noisy variables in PCA*
>
> In PCA, it is customary to scale columns of ***X*** so that in the scaled matrix, all of the columns have the same variance. This procedure means that the sum of the two components of the variance (signal and noise) is constant over all variables. It follows that for the weakest variables, having the smallest amount of signal, the noise variance is much larger than for the strong variables. This behavior is in severe conflict with the recommendations found in this work: the exaggerated noise in a few noisy variables will cause the small principal components to be undetectable in the analysis and will increase the noise in other principal components. The recommendation is clear: "do not auto-scale noisy variables in PCA modeling" […]
>
> .

Is this type of scaling also done in here? Please reflect and add some discussion on the issue of error model in PCA. Perhaps recommend the readers to note this for future analyses?

p.5.l.24. **:** While potentially simplifying the chemical interpretation of factors, rotating the variables in a direction where ions are more separated between the factors, does this not equally degrade the time series interpretability (loadings' time series get more similar)?

p.6.l.14.: Same question on EFA error model as for PCA above, as PCA and EFA are very similar methods. Please discuss error weighting and accounting signal-to-noise (Paatero and Hopke, 2003).

p.6.l.6.: Does the limit of 0.3 apply to AMS data as well? This means any m/z signal explaining under 0.3 ug/m3 was set to zero? How many variables does this affect – I would imagine it is a very large fraction? How does this truncation affect mass (signal) conservation in data?

p.7.l.15 On the error model of st.dev / sqrt(n): as written in the text, st.dev reflects the changes in the concentrations of compounds in the chamber, during the experiment, and not measurement uncertainty (or counting statistics error of the instrument) in a repetition measurement? How does dividing it by (square root of) data series length make it a more relevant error metric? Please explain. Also: Why not use the standard error model computed by the AMS analysis software (Squirrel/Pika; e.g. Ulbrich et al., 2009)?

p.7.l.21. (and in the supplementary referred): In calculation of the signal-to-noise-ratio (SNR), you identified weak and bad variables (defined by their low SNR). However in contrast to usual practice (in PMF), you do not down-weight these signals. Especially since you are using a custom (not thoroughly validated) error model, I would listen by the advice, Paatero and Hopke (2003):

> *"Regarding weak/bad variables, the main result of this work is that even a small amount of overweighting is quite harmful and should be avoided. In contrast, moderate downweighting, by a factor of 2 or 3, never hurts much and sometimes is useful. Thus, it is recommended to routinely downweight all weak variables by a factor of 2 or 3. This practice will act as insurance and protect against occasions when the error level of some variables has been underestimated resulting in a risk of overweighting such variables. Regarding bad variables (where hardly any signal is visible from the noise), the recommendation is that such variables be entirely omitted from the model."*

The reasoning for not down-weighting only states that low SNR data has high noise-to-signal rate, which is only stating the evident, and not very useful. Also, the small number of variables is not really a good excuse to deviate from the practice (without at least some sensitivity analysis). While I agree the overall error from not doing this properly is likely smallish (for PMF), I recommend you acknowledge the issue and cite this "best practice", the Paatero and Hopke (2003) recommendation, for the readers – not to proliferate a deficient data pre-processing practice.

p.7.l.26., also Figure S4 (should be: "S7"?). The standard AMS error model is composed of a minimum (Gaussian) error (Ulbrich et al., 2009), related to electrical background noise of the instrument (background at zero signal) plus a counting statistics uncertainty that follows the Poisson distribution (error is proportional to the square root of signal intensity; Allan et al., 2003). This model seems different from the ones used in this paper. Please comment on this, and why you chose not to directly use the approach of Allan et al and Ulbrich et al. (2009), readily available from the AMS analysis software.

From the plot S4 (should be "S7"?) "PMF error schemes" it seems the constant error scheme ("Static error") overestimates error, especially for low concentrations, and the counting statistics (Poisson) error seems negligible (underestimated), even for the most abundant aerosol ion, at m/z 44, at highest concentrations. Please double check and report the error calculations here against what you get from the "Squirrel" AMS analysis software.

Why was the "signal following error" not used/tested for the AMS, since it seems to work for PTR-MS?

p.8. l.6.: How much does the Q (or Q/Q exp ratio) increase with these fpeak values?

p.9.l.10.: How do you interpret the negative loadings in EFA and PCA? Why were these solutions deemed physically sound, if they feature negative mass loadings?

p.10.l.5.: *"Depending on the aim of the study 5 and the type of the data, this property of cluster analysis may be considered either as an advantage or disadvantage."*. This is an important point in this paper overall, as is the difference between hard vs soft divisions of variables. Please elaborate, and e.g. give the reader some examples of data analysis of objectives where cluster analysis would be at an advantage or disadvantage.

p.10., **Section 3.3**.: As you state, interpretability is key. Chemical interpretability is discussed in a concise way. However, in addition to chemical interpretation, looking at loading time series is an important indicator. Do the time series reflect the kind reaction kinetics (in experiment chamber) and take place in reasonable timescales? This relates to e.g. Fig.13.

Comparing e.g. the PMF results in Figs. 13, S26, S27 – only the rank 3 Poisson-model (Fig S26-c) solutions' loadings (and factor 3 in Fig 13d) behave in a realistic way. The others anti-correlate highly, usually signaling they are over-resolved (unrealistically split) and usually then less components should be used. See for example f1 andf2 in Fig 13-a: correlation is undoubtedly close to -1 and the dynamics do not make sense – this simply seems like a bad factorization solution.

p.13, Section 4.: the factor profile figures (Figs 3 through 7) are really difficult to read this way! Please separate each factor to its own sub-plot, similar to Figure 12-b. Maybe put the fractional plots to supplementary material?

Figure 1, Figure 2. Please highlight in these figure factor numbers (e.g. a larger dot?) that were selected according to the evaluation metrics.

In all figures, for quick reference, please state if it is gas or particle (or AMS vs PTR-MS) data.

p.14. Section 4.1.2.: The difficulty to interpret PCA (negative) highlights that PCA separates principal components of variability, whether positive or negative, which can not [necessarily] directly be interpreted as physical [concentration] components of the system – the variability could be equally connected to losses of signal due to it chemically reacting away. Again, this ties to the objective of the analysis and should be discussed more clearly.

p.17.l.25. *"This is caused by the used error scheme, where errors are larger for the fast changes in the data (Fig. S4b)."* I am very confused. Is this "feature" intentional? This also relates to the question of p.7.l.15. Is the variability of ion concentration in chamber is indeed used as a metric for measurement uncertainty (even when scaled by sqrt(n)?. This seems a peculiar, to say the least, a very un-orthodox way to do error modeling in PMF. It could explain why PMF does surprisingly poorly compared to the other models for AMS data.

Importantly, please include a comparison of your error models versus "the standard" error model in-build in AMS data analysis software(s) (Squirrel / Pika / PET) and list the differences between the standard practice (see e.g. Allan et al., 2003; Ulbrich et al., 2009; Zhang et al., 2011) and your data pre-processing procedure.

p.21., Section 4.2.4.: I would have inspected the 2 factor solution, is it had percentage-wise the largest decrease in Q/Qexp. Usually it is also good practice to start from lower number that you can certainly interpret and continue to higher numbers. This could explain also why EFA and PAM suggest 2 factors (clusters). Please show these 2 factor (cluster) solutions (in the supplementary material).

The analysis on the error scheme issues seems correct.

I have to disagree on the interpretability of the PMF solutions presented here. This solution seems a mathematical one rather than physically meaningful. See also comment to **p.10.**, **Section 3.3**., Figure 13 etc. The time series indicate over-splitting of factors (extreme anti-correlation of f1,f2,f3) and the most of spectral profiles have extreme positive correlation (extreme high similarity by visual inspection). I don't really see many interpretable features in Figure 13. Mainly the HOA spectrum in Fig 13b, LV(f1,f2,f3)/SV(f4) split in 13 c&d.

Technical corrections:

p.3.l.4.: PTR generally refers to the ionization reaction and not the instrument, Please use PTR-MS or similar, common acronym.

p.7.l.30. t_s here is sampling time per m/z channel, when scanning (Allan et al., 2003), not to be confused with the total sampling time of the instrument, so please add this clarification.

p.21.l.13.: Please add a reference to m/z 57 and 59 link to HOA.

Supplementary material: please check figure numbering is in numerical order. Similarly to main part figures, please add if data is PTR—MS or AMS.

references

Allan, J.D., J.L. Jimenez, H. Coe, K.N. Bower, P.I. Williams, and D.R. Worsnop, Quantitative Sampling Using an Aerodyne Aerosol Mass Spectrometer. Part 1: Techniques of Data Interpretation and Error Analysis, *Journal of Geophysical Research- Atmospheres*, Vol. 108, No. D3, 4090, doi:10.1029/2002JD002358, 2003. (NB! Includes corrigendum)

I.M. Ulbrich, M.R. Canagaratna, Q. Zhang, D.R. Worsnop, and J.L. Jimenez. Interpretation of Organic Components from Positive Matrix Factorization of Aerosol Mass Spectrometric Data. *Atmospheric Chemistry and Physics* , 9(9), 2891-2918, 2009.

P. Paatero and P. K. Hopke, "Discarding or Downweighting High-Noise Variables in Factor Analytic Models," Analytical Chimica Acta. Vol. 490, No. 1-2, 2003, pp. 277-289. http://dx.doi.org/10.1016/S0003-2670(02)01643-4

Q. Zhang, J.L. Jimenez, M.R. Canagaratna, I.M. Ulbrich, S.N. Ng, D.R. Worsnop, and Y. Sun. Understanding Atmospheric Organic Aerosols via Factor Analysis of Aerosol Mass Spectrometry: a Review. Analytical and Bioanalytical Chemistry, 401, 3045-3067, DOI:10.1007/s00216-011-5355-y, 2011.