

Response to reviewer 1

We thank the Reviewer for carefully reviewing our manuscript and providing insightful comments. Below we address each comment point by point. For clarity we mark the reviewer comment in **blue**, our answers in **black**, and changes to the manuscript in **red**. Page and line numbers (in **black**) in our replies refer to the clean revised manuscript (without tracked changes), and **green** line/page numbers refer to this response.

Summary:

10 This work accomplishes a cross-comparison of several data-reduction analysis techniques. The analysis is performed on a chamber experiment. Car exhaust was directly sampled into an environmental chamber. A-pinene was also added to the chamber. The mixed car exhaust/pinene was then aged via OH-initiated photooxidation. Two instruments were used, a PTR-ToF MS and an AMS. Data were then analyzed using principal component analysis, positive matrix factorization, 15 exploratory factor analysis, clustering, and non-negative matrix factorization. The resulting simplifications are compared in terms of the ability to reconstruct the original data set (residual), number of factors/groups required to explain data variability, time-series behavior of the factors, and chemical composition of the factors. The authors find that the preferred number of factors is roughly similar regardless of technique. Some factors (or groupings) are generally consistent (in 20 terms of time-series behavior and composition) regardless of technique. Some techniques, particularly PCA and EFA were found to be difficult to interpret and not as useful. The clustering method was not useful for AMS data.

Major comments:

25 1) The pre-light mixing period is a substantial fraction of the whole experiment- is this interesting? The inclusion of this time period seems to have significant impact on the algorithm results.

30 One of our aims was to investigate the robustness of the different SDRTs with a data set having large changes in the concentration. The mixing period was important to include in the analysis as the step change when the photochemistry is started in the chamber is a good test scenario because we know the reason for the change and the exact time. Also, the direction of change is known. Please see our reply to specific comment (“page 7 line 16”) in **page 4**.

35 2) Mostly this paper seems to show that various different algorithms distinguish similarly between primary and secondary VOCs. Is this a useful reduction of chamber data, to group all secondary VOCs into one or two blocs? From Table 1 it seems that the two matrix factorization techniques result in perhaps three oxidized factors, but this is hardly discussed in the text. It is not clear to me if the three oxidized factors are consistent between the two techniques. It would be helpful to have more interpretation of these factors, especially if it were supported by a more detailed 40 connection to the chemistry of the system. NMF also seems to result in some mixing of primary and secondary emissions (FN4) which is probably unphysical.

45 We want to point out that the PTR-MS mainly provides information on the precursors and not the heavily oxygenated compounds created in their oxidation. Thus, the data set is far from giving a full picture of the chemistry in the chamber during photo oxidation (see also our reply to specific comment “Page 24 line 13-14” in **page 10**). Separating primary emissions and grouping the reaction products into three groups without adding any additional information about the ongoing reactions is the first step for understanding the chemical processes in these experiments. For a more detailed 50 study, gas phase data from different mass spectra that cover a wider range of oxidation products

(e.g. PTR-MS and I CIMS) need to be combined. But before this can be attempted, we wanted to investigate the performance of the available SDTRs which to our knowledge has not been performed in such detail yet. However, we added the factor contributions as separate spectra to the SI section S5 to clarify the similarity of the acquired results of the different SDRTs used (Figures S21 and S22) and added a few more comments there in addition to the more “mathematical” comparison of the factors by contrast angle.

S5, page 13, lines 16-20

[...] Figure S21 and S22 show the factor contributions in separate panels for each factor for the main results from gas phase data (PTR-MS). Despite the fundamental differences between these methods, the relative factor contributions clearly show similarity. Even PAM, that is relatively different when compared to other SDRTs as it only assigns one m/z into one factor, the same ions are enhanced when compared to other SDRTs. [...]

3) The way the mass spectra (chemical composition of groupings) are presented is very difficult to interpret and compare. From Table 1 the authors have assigned a consistent identity to factors resulting from each technique. Can you show a direct comparison of “Factor 2” for example, perhaps by plotting one mass spectra against the other so that it is easy to see which VOCs are similarly enhanced, and which may be different?

We now added figures with the factor mass spectra in separate panels in the SI material section S5 for easier comparison (Figures S21 and S22). In addition, the contrast angle -procedure introduced in SI section S5 shows a direct mathematical way to compare the factor spectra between the different SDRTs.

Specific comments:

Page 3, line 3-4: I disagree with this statement, “PMF was originally developed for field measurement datasets where real changes in factors are expected to be much slower than e.g. the noise in the data.” In the ambient environment, VOC composition and concentrations can actually change very quickly (on the order of seconds), especially when plumes of highly-concentrated primary emissions are intercepted. The abrupt changes in conditions during a chamber experiment therefore do not present a special challenge to PMF, compared to ambient measurements.

We agree with the reviewer that also in ambient conditions, the changes can be fast depending on the environment. Hence our statement was misleading. We wanted to highlight here with our statement, that most of the PMF applications, so far, have included ambient measurement data. We have now reworded our claim into:

page 3, lines 5-8

[...] The special conditions in lab experiments (sharp change at the beginning of experiments, e.g., switching on UV-lights) present an additional test scenario, as PMF has been mostly used for field measurement data sets where the main focus is often in the long-term trends and real changes in factors are often expected to be more subtle, than e.g. the variations in the noise in the data. [...]

We also adjusted section 4.1.5 accordingly:

page 19, lines 17-24

[...] This is caused by the used error scheme, where errors are larger for the fast changes in the data (Fig. S4b). In ambient data not measured at instant proximity of strong emission sources, for which PMF is often used, this type of error is beneficial as there the fast changes are more likely to be noise or instrument malfunctions (excluding, for example, sudden primary emission plumes), and

5 we are more interested of the long-term changes instead. For laboratory data, where large changes are often caused by rapid changes in actual experimental conditions, e.g. due to injecting α -pinene or turning the UV lights on, the static type of error is most likely preferable. Usage of the static error scheme helps to avoid overcorrecting intentional (large) changes in experimental conditions and confusing them with real variation taking place during the experiment and typically being much less pronounced. [...]

10 Page 3, lines 26-29: A few more details are needed here (instead of in the supplement):
What was the typical concentration of total VOC (or of a few key VOC e.g. aromatics)?
What concentration of α -pinene was added?
What was the VOC-NO_x ratio, and how was it adjusted?
Why were these specific concentrations of vehicle exhaust VOC and α -pinene chosen?
Were vehicle exhaust and pinene added just at the beginning of the experiment, or were they continuously injected?
15 Was the chamber continuously refilled (and with clean air or with fresh emissions?) to replace air taken by the mass spectrometers, or did the volume of the chamber decrease over time?

20 We aimed to keep this description as brief as possible as the prime focus of this manuscript was the comparison of the SRDTs. Also, the experimental conditions for the whole measurement campaign are described and interpreted in detail in Kari et al. (2019). However, we adjusted section 2 to include the requested information, and also the mass resolution details requested by reviewer#2:

page 3 line 31 – page 4 line 4
25 [...] The exhaust was diluted using a two-stage dilution system and fed into the 29 m³ collapsible environmental PTFE chamber ILMARI (Leskinen et al., 2015). For the experiment investigated in this study, α -pinene (~ 1 μ L, corresponding to 5 ppbV) was injected into the chamber to resemble biogenic VOCs in typical suburban areas in Finland. Atmospherically relevant conditions were simulated by adding O₃ to convert extra NO from vehicle emissions to NO₂ and adding more NO₂ to the chamber if needed. With these additions, atmospherically relevant VOC-to-NO_x (~ 7.4
30 ppbC/ppb) and NO₂-to-NO ratios were achieved to resemble the typical observed level in suburban areas (National Research Council, 1991). [...]

page 4, lines 10-22
35 [...] Volatile organic compounds (VOCs) in the gas phase were monitored with a proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF MS 8000, Ionicon Analytik, Austria, hereafter referred to as PTR-MS). Typical concentration for few example VOCs in the mid-way of the experiment were 2 μ m/m³ for toluene, 0.2 μ m/m³ for TMB (trimethylbenzene) and 1.7 μ m/m³ for C₄H₄O₃. Detailed setup, calibration procedure and data analysis of the used high resolution PTR-MS have been explicitly presented in Kari et al., (2019b). In the campaign, the high mass resolution of the instrument (>5000) enabled the determination of the elemental compositions of measured
40 VOCs. The instrumental setting was intended to minimize the fragmentation of most compounds, so the quantitation of the VOCs was possible. The chemical composition of the particle phase of the formed SOA was monitored with a soot particle high resolution aerosol mass spectrometer (SP-AMS, Aerodyne Research Inc., USA, hereafter referred to only as AMS, Onasch et al., 2012). In brief, the SP-AMS was operated at 5 min saving cycles, switching between the electron ionization
45 (EI) mode and SP mode. In EI mode, the V-mode mass spectra were processed to determine the aerosol mass concentration and size distribution. The mass resolution in this mode reaches ~2000. The SP-mode mass spectra were used to obtain black carbon concentration. As the used chamber

was a collapsible bag, the volume of the chamber decreased over time due to the air taken by the instruments. [...]

5 Section 3.1.2: EFA seems very similar to PMF. Could you please explain the major relevant difference(s) between EFA and PMF, and how they would affect the resulting dimensionality reduction?

10 The main difference between EFA and PMF is that in EFA, the factorization algorithm is applied to the correlation matrix constructed from the data, whereas in PMF, the original data matrix is factorized. This means, that in EFA, the created factors include variables, that have stronger correlation compared to variables from other factors.

15 In PMF, the error matrix is used as a tool to downweigh noisy (or weak) signals. This can also be used to put emphases on certain parts of the data set (e.g. the part with most change in signal). EFA, however, has other options to discard noisy or insignificant signals (see last paragraph in section 3.1.1).

20 We have now added a summary section (new section 4.4.) to address and discuss the similarities and differences between the presented SDRTs as requested here and by reviewer#2.

Page 7 line 16 (and elsewhere): At multiple points in the manuscript it is mentioned that the rapid changes associated with lights-on cause problems when implementing the various dimensionality reduction techniques. Would it make more sense to exclude data prior to $t=0$? Was there a reason this was not done?

25 For a detailed analysis of such a chamber experiment, the data before the onset of photo chemistry would indeed be omitted. But the main focus of our study was to compare the performance of the different SDRTs with different types of mass spectra data. As the reviewer points out above, rapid changes in composition can also occur in the atmosphere. Thus, it is important to test the response of all SDTRs to such a change. The induced step change when photochemistry is started in the chamber is a good test scenario as we know what caused the change and the exact time. Also, the direction of change is known (precursor getting consumed and products formed).

We added a few sentences to the manuscript to further justify our selection of the data:

35 section 1, page 3, lines 24-25

[...] Further, we examine the performance of the SDTRs when the data includes large and rapid changes in the composition. [...]

section 2, page 5, lines 2-4

40 [...] In addition, as the main focus of our study was to compare the performance of the different SDRTs with different types of mass spectra, instead of detailed analysis of the chamber experiment, we have also included the pre-mixing period during the α -pinene injection (i.e., $t < 0$) into our analysis. [...]

45 Page 10 lines 28-31: Since Figure 1 relies on a comparison of BIC and SRMR, it would be helpful here to provide more detail on how these two metrics are calculated, what the relevant differences are, and why one may be preferred over the other. What was the purpose of calculating both metrics and why were these particular metrics chosen?

These particular metrics were selected, as they measure slightly different properties of the model. BIC is a comparative measure of fit balancing between increased likelihood of the model by adding parameters and a penalty term for number of parameters. The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. It is a positively biased measure and that bias is greater for small N and for low df (degrees of freedom) studies.

The definitions and equations of these metrics can be now found from the SI material (new section S3.2), and we added a few sentences to manuscript section 3.3, as requested (page 12, lines 12-15):

[...] These metrics measure slightly different properties of the model. BIC is a comparative measure of the fit, balancing between increased likelihood of the model and a penalty term for number of parameters. The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. See S3.2 for more details. [...]

Equations 11 and 12: There are several errors in these equations which are likely a copying error from Brunet et al. 2004. The authors should check that the actual implementation was done correctly. The equations should read:

$$H_{au} \leftarrow H_{au} \frac{\sum_i W_{ia} X_{iu} / (WH)_{iu}}{\sum_k W_{ka}}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_u H_{au} X_{iu} / (WH)_{iu}}{\sum_v H_{av}}$$

I also suggest here to use “k” as the row index for W (in the denominator term), to avoid confusion with “p” being the factor rank, and for consistency with Lee and Seung, 2001.

We thank the reviewer to pointing this out. But while the formula was incorrect in the manuscript text, all calculations we performed with the correct set of equations. Thus, none of the results for NMF need to be changed. The equation is now corrected, and the index “p” is changed to “k” as suggested. We added a sentence before equation (11) to clarify the connection between k and p:

page 9, line 20

[...] The value of k is equivalent to the selected factorization rank p . [...]

Page 11 lines 29-33: Given that the update functions (11) and (12) are derived from the divergence cost function $D(X||WH)$ (Lee and Seung, 2001, Eq. 3), I suggest that this cost function is monitored as a function of p , analogously to $Q/Q_{exp}(p)$ for PMF. The termination condition for NMF wasn't described in Section 3.1.4, but presumably it is not dependent on p ; if this is the case then the divergence of the end solution can be compared for each value of p . The residual sum of squares is not an appropriate metric, as this was not the cost function used for the NMF implementation.

The update functions presented in the manuscript are indeed derived from the Eq. 3 presented in Lee and Seung (2001). We have now modified the manuscript and use the value of the cost function in the last iteration for each p instead of the RSS, as suggested.

We modified the last paragraph of section 3.3, corresponding figures (Figure 2 and Figure 11), and the results (sections 4.1.4 and 4.2.3) accordingly:

section 3.3

page xx, lines xx-xx

[...] In addition, we investigated the cost function that approximates the quality of factorization as a function of the factorization rank p . For the brunet-algorithm that we applied in this study, this cost function is the divergence between data matrix X and the approximation WH (see Eg. (3) in Lee and Seung (2001)). [...]

5

section 4.1.4
page xx, lines xx-xx

[...] Figure 2a shows the divergence of the cost function $D(X||WH)$ and CCC for factorization ranks from 2 to 10 for NMF. The CCC has a first decrease in the values at the rank 4 and the $D(X||WH)$ shows an inflection point around the ranks 4-5. Figure 6 shows the factor time series and total contribution for the NMF with factorization rank 5. Five factors were selected, even though CCC suggest only 4 factors, as [...]

10

section 4.2.3
page xx, lines xx-xx

[...] The $D(X||WH)$ has an inflection point at factorization rank 4 and CCC shows the first decrease in the values with 4 factors, as shown in Fig. 11a. [...]

15

Page 12 line 2: What is meant by “not achieved only by change?”

20

The purpose of the resampling here was to approximate and reduce the uncertainty in the factorization, based only on one dataset. As the figures show, the results vary significantly, especially when fast changes are taking place. This points out that all methods are rather sensitive for only small changes in the data and thus this kind of sensitivity test is necessary.

25

In order to avoid confusion by wording, the first sentences of section 3.4 (page 13, lines 20-23) were reformulated to:

[...] When analyzing the datasets, we realized that all of the factorization methods in this study are sensitive to even small changes in the data. In order to cross-validate the calculated factorization and approximate the uncertainty in the factors, 20 resamples of the measurement data were created with bootstrap-type sampling (Efron and Tibshirani, 1986), i.e., sampling with replacement from the original data. [...]

30

Page 13 line 5: Why not compare the absolute value of the residual?

35

For PMF and NMF, which follow the matrix equation $X = WH + E$, where E is the residual matrix, this type of comparison is applied in by calculating the total residuals as a function of time (i.e. summing all variables for each time point in E), see e.g. page 24 lines 11-12 for AMS. However, as described in section 3.4 (see page 14, lines 16-25), this type of residual calculation is not possible for EFA and PCA, as they factorize the correlation matrix and do not create W and H in a similar manner (i.e. simultaneously). Therefore, we can calculate the residuals for EFA/PCA only by reconstructing the correlation matrix (these values are shown in the manuscript, e.g. for EFA in sect. 4.1.1, page 15, line 24), but not the actual data. Thus, the actual residual values between PMF/NMF and PCA/EFA are not comparable. In addition, in PAM, which uses the distance matrix in the “decomposition”, the situation is again very different.

40

45

Pages 14 and 15: For other researchers which would like to use this paper as a guide, it would be helpful to indicate the range of values that are acceptable. For example Page 14 line 4-5, what value of residual would be considered not acceptable? Page 14 line 32, what is the Kaiser limit and what

range of values are considered “close”? Page 15 line 11 are these considered large or small residuals?

5 As desirable as such absolute values would be, unfortunately it is not possible to set them in a general manner. Especially for PMF and NMF, there is a strong subjective element in the selection of p values and in the magnitude of acceptable residuals. It may depend on the type of data and the purpose of the analysis what is considered acceptable. The values may differ between analyzing a long ambient data set for the possible SOA sources or types (e.g. OOA vs HOA in an AMS data set) and studying a chamber data set trying to identify chemical processes. Please also note that most of these single value “goodness of fit” parameters are summarized over all observations and variables. 10 It is known that for PMF the overall Q values may be low while Q values for individual variables are not (interactive comment from Paatero to Yan et al., 2016).

15 The Kaiser limit/criterion refers to the ideal value of 1, and components having lower eigenvalues should be discarded (explained in sect. 3.3, page 12, lines 6-7). What is meant by being “rather close” to this limit, is that the relative decrease of the eigenvalues is getting smaller and smaller with increasing number of components. So even though we reach the Kaiser limit (our eigenvalues get smaller than 1) having as many as 9 components, it is not realistic to have this type of system with that many components. In addition, as mentioned in sect. 3.3 page 12, lines 17-20, this type of 20 tests for the number of components/factors are not meant to be taken as strict rules, but rather give the first insights into the possible dimensions of the investigated system. We determined the most suitable number of dimensions for each data after investigating multiple solutions from various SDRTs.

25 As explained in sect. 3.4 (page 14, lines 27-29) the theoretical maximum value for the mean and IQR of the absolute residual correlations is 1 (indicating in principle no reconstruction at all) and the minimum 0 (perfect reconstruction). For example, if we would have a correlation coefficient of 0.6, and the absolute residual correlation is 0.0116 (as given the mean value in the line referred for PCA), we over- or underestimate the correlation coefficient by 1.93 % $((0.0116/0.6)*100)$. Thus, 30 the smaller values we have, the better the reconstruction. Unfortunately, there are no general guidelines, and our main aim here was to compare if the reconstruction differs between EFA and PCA. To give the reader a more concrete way to understand the actual “size” of the errors for EFA/PCA in this study, we added the similar example to sect. 3.4:

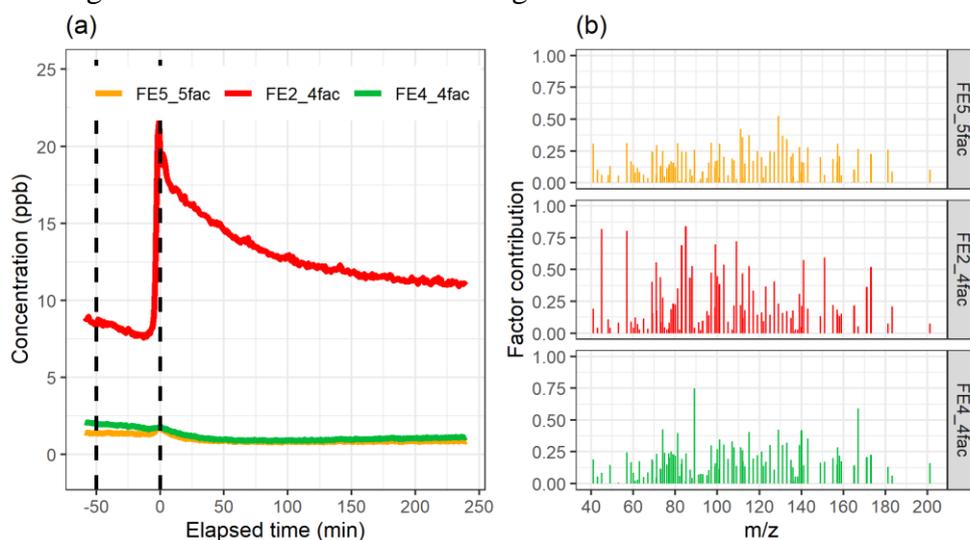
35 page xx, lines xx-xx

[...] For example, for a variable-pair having a correlation coefficient of 0.7, a mean absolute correlation residual of 0.02, and an IQR of 0.04, this would mean the model over- or underestimates the correlation by 2.86% $(= (0.02/0.7) * 100)$. An IQR of 0.15 would mean that 50 % of all variable-pairs with correlation of 0.7 are within 5.7% $(= (0.04/0.7) * 100)$ of the original value of 0.7. [...]

40 Page 14 line 1 and page 15 line 2: Can you show please how it is determined that the additional component is not a new component with different properties but rather a mixture of previous components?

45 Figure 1 below shows the related factors (FE5 from the 5-factor case and FE2&FE4 from 4-factor case) for EFA. From the time series we can see that the FE5 has a small peak at $t = 0$, suggesting it has picked up some properties from FE2. For example, m/z 89.06 in FE4 in the 4-factor case (Fig. 2b, contribution of 0.74 in FE4, others < 0.16), is divided more “equally” between the factors when using 5 factors (contribution of 0.17, 0.04, 0.11, 0.42, 0.26 for factors FE1-FE5), thus not being as

well separated. Similarly, some other m/z are more scattered between the factors in the 5-factor case, whereas with 4 factors they are more or less assigned to one, or possibly to 2, factor(s) only. This indicates the addition of the 5th factor does not introduce new properties but rather splits the existing ones into subsets. The reasoning for PCA is similar.



5

Figure 1 Factors 2 and 4 from the 4-factor case and factor 5 from the 5-factor case. (a) shows the time series and (b) the factor contribution.

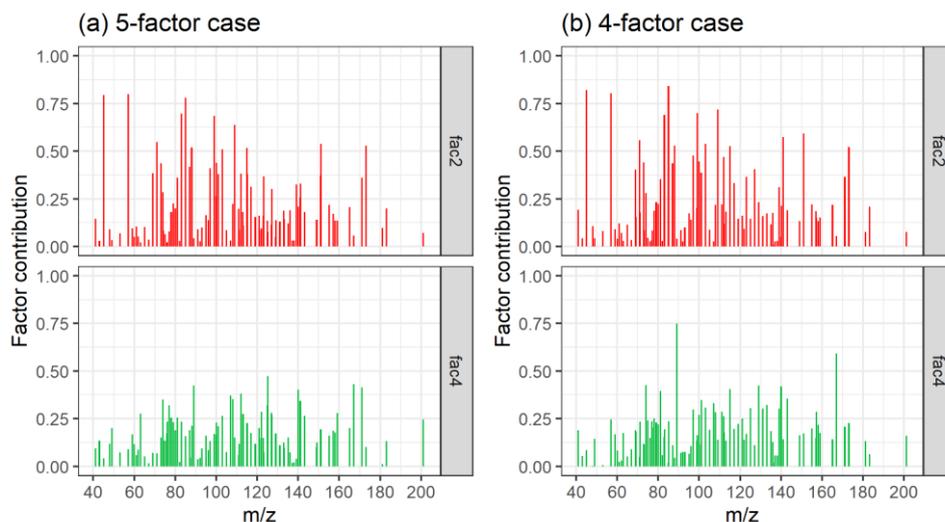


Figure 2 Comparison of the factor contributions for factors 2 and 4 from the (a) 5-factor case and (b) 4-factor case.

10

Page 17 Lines 16-17, 23-30: The signal following error is essentially introducing a smoothness constraint, which doesn't seem appropriate given that you know there are sharp changes due to experimental conditions. Is it recalculated for each data resampling? Why not resample the error along with the data? Is it possible to split the data into time periods whose start and stop are defined by sample injection, lights on, etc. so that the running standard deviation does not include these sharp changes? Additionally, lines 26-27: Ambient data often has fast changes that are due to real variability. This is one of the reasons why fast online techniques such as PTR-MS are used for ambient measurements, because they allow the observation of these changes.

15

20

One of our aims in this study was to see how PMF performs with very different error estimates. Thus, we introduced the static error, that gives equal weight for every time point for each variable in question, and the signal following error, which has larger errors for rapid changes. To split the

5 data set into sections requires a priori information, which is usually not available especially for ambient data sets. For our laboratory measurements, we naturally have this information, but as we specifically wanted an error estimate, that could possibly account for the large changes, we did not split the data for the error calculation. Instead, we also had the static error giving equal weights for each time point. We reworded the section 4.1.5 to clarify, please see reply to specific (“Page 3, line 3-4”) comment in [page 2](#).

10 For each data resampling we recalculated the error, thus the larger errors were in the “correct” places for each resample too.

15 [Page 22 line 20](#): This indicates that the error estimation should be revised, or that these compounds should be downweighted. PMF and NMF are extremely similar techniques with the crucial difference being only the inclusion of an error matrix, so it does not seem likely that the difference in performance is due to the size of the dataset. The extremely small values of NMF residuals also seem suspect. The authors should check that residuals for PMF and NMF were calculated in exactly the same way so as to enable the direct quantitative comparison.

20 One of our aims was to investigate how each of the SDRTs perform with datasets possibly including noisy background ions, thus we did not downweighed them. As we also do not have similar downweighing option for NMF, we omitted this procedure for PMF too to give each method an equal “starting point”. In addition, the error estimates we used did have relatively large errors for these signals (noisy, small concentration) and therefore already giving PMF the possibility to not to give so much weight for this type of signal in the first place. As the PMF and NMF results were agreeing nicely with PTR-MS data, we do believe the issue here is either the data size and/or the used brunet-NMF is indeed computationally more efficient/accurate, compared to the version of PMF used here, when handling small data. In addition, PMF performed without any issues for a combined dataset from the same measurement campaign (Kari et al., 2019), indicating the issue in PMF here is indeed related to the small size of the data set.

30 However, we do agree that to achieve the best chemical interpretation (which was not the main aim of this study) for the factors from PMF, the downweighing is advisable. To justify our decision, we now also applied PMF with the downweighing, and no change in the results was observed. We now state this in the manuscript, and in the SI. In addition, as requested by reviewer#2, we have now included the 2-factor PMF solution for particle phase data to the manuscript and moved the 4-factor solution to the SI. Please see the reply to comment “Figure 13” in [page 11](#).

We have added more justification for omitting the downweighing into section 3.1.3 and S2.3:

40 section 3.1.3, page 8 line 27 – page 9 line 2
[...]. In contrast to the suggested best practice (Paatero and Hopke, 2003), we did not downweighed any ions in our data sets. This approach was used in order to give each SDRT an equal starting point for the analysis, as e.g. for NMF or PCA similar downweighing is not possible, as we do not have any error estimates to calculate the signal-to-noise ratios in similar manner. However, to avoid misguiding the reader to omit recommended data pre-processing practice for PMF, we also tested PMF with downweighing. This, as expected, did not change our results significantly, but we acknowledge it should be indeed applied if aiming for a more detailed chemical interpretation of the PMF factors. [...]

section S2.3, page 5, lines 1-5

[...] However, no removal or downweighing of the variables was applied in the results presented in the main manuscript, as for the weak ions the error values were already in the range of the ion signal itself, and for the bad variables the error was usually way above the signal throughout the ion time series. In addition, the number of weak or bad ions for both AMS and PTR-MS data were rather small. To justify our decision, we did run the PMF analysis with the downweighing. This, as expected, did not change our results or interpretation, and thus those results are omitted. [...]

We re-checked the NMF and PMF residuals, and they both are calculated exactly the same way (i.e., multiplying the factorization matrices W and H (or G and F in PMF) to obtain the reconstructed data matrix, and then calculating the difference between the data and the reconstruction for each point in time). The very small residuals in NMF suggest the optimization with the brunet algorithm in NMF is performing relatively better than the PMF algorithm. However, better mathematical performance does not necessarily mean easier/better interpretation of the physical/chemical properties. Brunet-NMF was also selected, as the residuals were smallest when comparing to other available algorithms for NMF. Please note that there also exist other algorithms for PMF, depending on the interface that is used to apply PMF, and those might have different performance (see. e.g. Ramadan et al., 2003).

Page 24 line 13-14: Is this correct? I read this paper as well. I thought they had PTRMS and AMS. Additionally, PTR-TOF is a subset of TOF-CIMS, or?

By definition, every PTR-MS is a chemical ionization instrument. So, a PTR-TOF could be called a H_3O^+ TOF-CIMS. But one has to be careful as these labels are often used to differentiate between different instrument designs. The main difference lies in the design of the sample inlet and the actual ionization region. The PTR3 (used by Koss et al. 2020) has a much improved overall sensitivity over the earlier PTR-TOF due to changes to the electric field guiding the ions through the ionization region. Additionally, the sample inlet was re-designed to better transmit lower volatility compounds (i.e. mostly higher oxygenated) which were previously lost as they are too “sticky”. Riva et al. (2019) compared the performance of different chemical ionization instruments. The VOCUS PTR is here similar to the H_3O^+ PTR3. C10 compounds with 3 or more oxygen atoms are barely visible in a PTR-TOF but can be detected with a VOCUS-PTR (Figure 6 in Riva et al. 2019). The Aerodyne I CIMS (which has yet another design for the inlet and ion reaction region) mostly detects C10 compounds with 2 to 8 oxygen atoms. Thus, we can conclude that the PTR-TOF used in our study mostly detects the precursors and early oxidation products (with generally low oxygen content) while the instruments used by Koss et al. (PTR3 and I CIMS) observe these compounds and also a large fraction of later generation oxidation products and also highly oxygenated compounds (HOM).

To clarify these differences in the instruments we modified our sentence to:

page 27, lines 25-27

[...] In addition, Koss et al. (2020) used gas-phase data from I-CIMS and PTR3 with NH_4^+ as a reagent ion, which are more sensitive to later generation oxidation products compared to the PTR-MS which we have used here. [...]

Figures 3,4: In the plot caption or legend it would be helpful to have a brief description of the interpretation of each factor, e.g. “pinene”, “car exhaust”, “background”. Additionally the display in plot (b) of factor contribution as a function of m/z doesn’t add much to the paper; I wouldn’t expect the factor contribution to depend on m/z in any particularly meaningful way. Since (I believe) your

PTRMS has multiple peaks resolved at each nominal mass, showing a unit-mass stick spectrum here is also not especially meaningful. If you want to show mass spectra, I strongly suggest to break panel b into 4 separate spectra, one for each factor, so that they can be examined separately.

5 Note that in these figures we did indeed show the high-resolution data and not the UMR signals. The factor contribution plots (Figures from 3 to 7) were not shown to suggest any correlation of the factors to m/z. Rather, we looked for a way of displaying the factor/compound “compositions” in a way allowing direct comparison of the different methods. Normalizing the factor contribution for each ion was chosen to avoid the very different weighting methods in NMF and PMF, and to be able to compare how much each ion contributes to each factor between all SDRTs. The exact ion mass was used as a unique identifier for all ions as we did not have sum formulas or compound names assigned to all ions. We now added figures with the factor mass spectra in separate panels in the SI material for easier comparison (section S5, renamed to “Contrast angle and factor spectra comparison” Figures S20, S21), and refer to this in the beginning of section 4.1.6. We added description (those in table 1 & 2) to the figures in the manuscript (when applicable), as requested.

section 4.1.6, page 20, lines 1-2

[...] Table 1 summarizes the acquired results from different SDRTs for the gas phase composition data measured with PTR-MS, and Figs. S20 and S21 in SI section S5 show separate factor contributions for each of the SDRTs. [...]

Figure 9: Where does isoprene come from in this experiment? Is this more likely to be a hydrocarbon from vehicle exhaust? Cycloalkanes in fossil fuel are known to create PTR ions at C5H9+, see e.g. Yuan et al. Chem Rev. 2017 doi.org/10.1021/acs.chemrev.7b00325.

25 We used the label “isoprene” for the ion C5H9+ following general naming conventions in the PTR community. But the reviewer is correct in assuming that here the signal at C5H9+ is connected to vehicle emissions and not biogenic emissions. The factor/component contributions of this ion are very similar to trimethylbenzene (TMB). Also, in PAM the C5H9+ is assigned to the cluster containing among others furan and TMB. Even with the short-comings of the clustering approach this strongly indicates the main correlation of C5H9+. We changed the label in Figure 9 and Figure 10 to C5H9+ to avoid the association with biogenic emissions that the label “isoprene” may cause.

Figure 13: The factor time series for the most part do not look realistic. What physical process could lead to the non-smooth behavior and multiple maxima?

35 Following the suggestion by reviewer#2, we adopted a 2-factor solution from the PMF results to the manuscript instead of the 4-factor solution (which is now moved into SI material). Please see the revised manuscript sections 4.2.4, 4.2.5 and the figures referred there.

40 **Minor/Technical corrections:**

Page 2 line 22, “alike” -> “like”

Page 3 line 22, “was” -> “were”

Page 13 line 24, “described in section 0” -> “described in section 3”

Page 15 line 19, “gab” -> “gap”

45 Page 19 line 27, “much” -> “many”

The suggested corrections were applied to the manuscript

REFERENCES

- Kari, E., et al.: Dual effect of anthropogenic emissions on the formation of biogenic SOA, *Atmos. Chem. Phys.*, 19, 15651-15671, 2019
- Koss, A. R., et al.: Dimensionality-reduction techniques for complex mass spectrometric datasets: application to laboratory atmospheric organic oxidation experiments, *Atmos. Chem. Phys.*, 20, 1021-1041, 2020.
- 5 Paatero, P.: Interactive comment on "Source characterization of Highly Oxidized Multifunctional Compounds in a Boreal Forest Environment using Positive Matrix Factorization" by Chao Yan et al., *Atmos. Chem. Phys. Discuss.*, 2016
- Riva, M., et al.: Evaluating the performance of five different chemical ionization techniques for detecting gaseous oxygenated organic species, *Atmos. Meas. Tech.*, 12, 2403–2421, 2019
- 10 Lee, D. D., and Seung, H. S.: Algorithms for non-negative matrix factorization, *Adv. Neural. Inf. Process. Syst.*, 13, 556-562, 2001.
- Ramadan, Z., et al.: Comparison of Positive Matrix Factorization and Multilinear Engine for the source apportionment of particulate pollutants, *Chemometr. Intell. Lab*, 666, 15-28, 2003.

Response to reviewer 2, Mikko Äijälä

We thank Dr. Äijälä for carefully reviewing our manuscript and providing very useful comments and suggestions, which improved our manuscript. Below we address each comment point by point.

5 For clarity we have marked the reviewer comment in **blue**, our answers in **black**, and changes to the manuscript in **red**. Page and line numbers (**in black**) in our replies refer to the clean revised manuscript (without tracked changes), and **green** line/page numbers refer to this response.

Summary

10 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 + α -pinene aging experiment in a reaction chamber. The data, collected by a PTR-MS and an AMS, was factorized and clustered using 5
15 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
20 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
25 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
30 after addressing the following comments.

Major comments

The authors generally do a good job of introducing and describing their methods, but I would call
35 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
40 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.

45 We have now summarized and added discussion of the used SDRTs to a new section in the manuscript (**section 4.4.**), where we aimed to discuss the objectives and purposes of these techniques in more concise manner than mentioning these properties throughout the manuscript.

4.4 Summary of the SDRTs used in this study

The methods tested in this study have many similarities and many, fundamental or computational, differences between them. Though in literature, they are many times applied to similar problems. In this section we will summarize some of these properties.

5 EFA has a fundamental difference to the other methods as it is by definition a measurement model of a latent variable, i.e. the factor, (Osborne, 2014) whereas the other methods are basically describing the measured data with linear combinations of measured variables. The latent variables in EFA, i.e. the factors, cannot be directly measured, but instead, they are seen through the relationships they initiate in a set of Y variables, which are measured. In the other methods, in turn, the factors, components or clusters are calculated directly from the measured variables Y (Rencher and Christensen, 2012; Osborne, 2014).
10

The approach to data reduction in PCA is to create one or more summarizing variables from a larger set of measured variables by retaining as much as possible of the variation present in the original data set (e.g. Jolliffe, 2002). This is done by using a linear combination of a set of variables. The created summarizing variables are called components. The main idea of the PCA is to figure out how to optimize this process: the optimal number of components, the optimal choice of measured variables for each component, and the optimal weights when calculating the component scores.
15

The objective of cluster analysis is to divide the observations into homogeneous and distinct groups (Rencher and Christensen, 2012). Cluster analysis is a method where the aim is to discover unknown groups in the data, which are not known in advance. The goal of the clustering algorithm is to partition the observations into homogeneous groups by using some measure of similarity (or dissimilarity), such that the within-group similarities are large compared to the between-group similarities. The choice of the similarity measure can have a large effect on the result. One property of cluster analysis is that it will always calculate clusters, even if there is no strong similarity present between the variables in the data (Wu, 2012). This should be noted when interpreting the results, especially if the user has no a priori information about the number of clusters.
20
25

NMF and PMF provide an alternative approach to the decomposition assuming that the data and the components are non-negative (Paatero and Tapper, 1994; Lee and Seung, 1999). Thus, all the features learned via NMF and PMF are additive; that is, they are adding together strictly positive features. PCA and EFA tend to group both positively correlated and negatively correlated components together, as they are only looking for the correlations of variables (except SVD-PCA, which can be applied to the data matrix directly). On the other hand, NMF and PMF, by constricting W and H to positive values, find patterns with the same direction of correlation. Thus, NMF and PMF work well for modelling non-negative data with positive correlations. However, if the interest is not only in the positive effects, then PCA and EFA can provide more information of for the investigated system. Cluster analysis is suitable for classifying observations based on certain criteria. The researcher can measure certain aspects of a group and divide them into specific categories using cluster analysis. However, this method is not suitable for data with variables which should show contributions from multiple factors/components (e.g. strongly fragmented signals in AMS data).
30
35

40 Factorization methods, including those used in this paper, operate on the fundamental assumption that the factor profiles (here factor mass spectra) are constant over the investigated period. Often, this has been interpreted in a way that chemical processes occurring in a chamber experiment or the atmosphere violate this assumption. However, this interpretation is based on a too narrow definition of what a factor represents. A factor can be seen as a direct (emission) source of compounds which

changes its contribution to the whole signal (e.g. primary emissions from biomass burning as a fire develops and then dies). But a factor can also be interpreted as a group of compounds showing the same temporal behavior. If this group is released together as an emission or if the compounds are formed in the same ratio by some chemical process should not matter. In the latter case, it is important how wide the group is selected, i.e. if we group products of processes together for which the contribution changes with time. This means, choosing the optimal number of factors becomes even more important when chemical processes are occurring. EFA and PCA account for the chemistry happening in chamber measurements with negative loadings, as described above. Same factor can contain educts and products of a chemical process (e.g. oxidation), with the difference that their loadings are negative and positive, respectively.

Taking account of everything above, the most important thing to consider when selecting the SDRT is the interpretation. What are the features the researcher wants to deduct from the data, what are the properties of the data and how the data can answer the research questions. As we have shown in the results section, the methods provide quite similar reconstructions of the time series, but the interpretation of the steps leading to these are quite different. For example, comparing EFA factor FE4 in Fig. 3a and PMF factor FP4 in Fig. 7a reconstructions, they seem to show the same procedure but the first one includes both positive and negative effects and the second one consists only of positive effects.

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.

In our manuscript we did use the standard AMS error model for PMF, and we understand that this was clearly not emphasized enough in the manuscript as our error description was misunderstood. The main point using the standard AMS error was indeed to show how “reference” error works for PMF when compared to others. We have addressed this issue in detail in the specific comments (see our reply to comment “p.7.1.26” in [page 10](#)) and modified the manuscript accordingly. We have also now named the “Poisson style error” to “standard AMS error” to emphasize this.

We have also clarified the data preprocessing steps for EFA and PCA as suggested, see our replies to specific comments “p.5.1.30” and “p.6.1.14” in [pages 6 and 8](#), respectively.

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.

The assumption of constant factor profiles is indeed a very important one for PMF (and NMF). The idea that chemistry in the atmosphere/chamber violates this assumption is however a misconception stemming from a too narrow definition of what a factor represents. A factor can be seen as a direct (emission) source of compounds which changes its contribution to the whole signal (e.g. primary emissions from biomass burning as a fire develops and then dies). But a factor can also be interpreted as a group of compounds showing the same temporal behavior. If this group is released

together as an emission or if the compounds are formed in the same ratio by some chemical process should not matter. In the latter case, it is important how wide the group is selected, i.e. if we group products of processes together for which the contribution changes with time.

5 As an example, consider a data set of gas phase measurements in the atmosphere. There is a mix of precursors and let there be OH and O₃ chemistry during the day and NO₃ and O₃ chemistry at night. All of these reactions can form Highly Oxygenated Material (HOM) among other compounds. The ratio of the reaction pathways (and thus the observed HOM composition) will depend on the ambient conditions (e.g. O₃ and NO_x concentration, irradiation). If we do not constrain the PMF analysis, we will find more than one HOM-type factor (i.e. a daytime and a night-time factor). Each of these factors then has a constant profile, only the contribution to the data set changes with time, i.e., the time series of the night-time factor will be close to 0 during the day and vice versa for the day-time factor. As there is no strict criterion for the correct number of factors, it is important that the scientist analyzing such a data set compares solutions with different number of factors and uses additional information (such as other measurements).

10 We also could perform a “rolling window” style PMF (looking at a short period of the data set at a time (e.g. an hour) and then shifting this window by e.g. 10 min). If we set the number of factors to only find 1 HOM type factor in the whole data set, this factor would have a different factor profile during the day and at night. But this behavior was created by the choice of the scientist to have a certain number of factors.

15 In the analysis presented in our manuscript, we did not use any such assumption about the type of factors we were expecting or even their number. Rather we inspected the solutions for multiple factor numbers and chose the number of factors based on the presented criteria and the interpretability, i.e., we first ran the analysis and then interpreted the results. We have now stated the assumptions for constant factor profiles in the new section 4.4 with further discussion. See reply to the first major comment in [pages 1 and 2](#).

Specific comments:

20 **p.2.1.2.:** “...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!

25 Indeed, PMF does not make any (mathematical) assumptions of the order of measurement points or rates of change. However, most of the PMF applications, so far, have been with ambient measurement data where, for most part of the data, the instrument noise is indeed more rapidly changing than the actual observed concentrations, which must be also assumed e.g. in the study by Ulbrich et al. (2009), where the AMS error scheme was introduced. We have now reworded the referred sentence to clarify that:

30 page 3, lines 5-8

35 [...] The special conditions in lab experiments (sharp change at the beginning of experiments, e.g., switching on UV-lights) present an additional test scenario, as PMF has been mostly used for field measurement data sets where the main focus is often in the long-term trends and real changes in factors are often expected to be more subtle, than e.g. the variations in the noise in the data. [...]

40 We also adjusted section 4.1.5 accordingly, please see reply to comment (“p.17.1.25”) in [page 15](#).

45 **p.2.1.14.** “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.

5

We have added a **new section 4.4** to the manuscript to include more discussion of the fundamental differences and objectives the various SDRTs might have. Please see our answer to the major comment in the beginning of this reply.

10 **p.2.1.16:** “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.

15 Please see our reply to specific comment “p.10.1.5.” in **page 13** for added examples. We modified the referred sentence to

section 1, page 3, lines 19-21

20 [...] On the other hand, clustering might be more suitable for a more simplified or preliminary approach (as it is computationally less demanding), or when the chemical compounds in the data are already known or if strict division between variables is preferred. [...]

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.

25

Both of these instruments are extensively described in Kari et al. (2019) and in the references given there. However, we added the requested details of the mass analyzer resolutions shortly to the manuscript along with the other details requested by reviewer#1;

30 section 2, page 4 lines 13-22

35 [...] Detailed setup, calibration procedure and data analysis of the used high resolution PTR-MS have been explicitly presented in Kari et al., (2019b). In the campaign, the high mass resolution of the instrument (>5000) enabled the determination of the elemental compositions of measured VOCs. The instrumental setting was intended to minimize the fragmentation of most compounds, so the quantitation of the VOCs was possible. The chemical composition of the particle phase of the formed SOA was monitored with a soot particle high resolution aerosol mass spectrometer (SP-AMS, Aerodyne Research Inc., USA, hereafter referred to only as AMS, Onasch et al., 2012). In brief, the SP-AMS was operated at 5 min saving cycles, switching between the electron ionization (EI) mode and SP mode. In EI mode, the V-mode mass spectra were processed to determine the aerosol mass concentration and size distribution. The mass resolution in this mode reaches ~2000. The SP-mode mass spectra were used to obtain black carbon concentration. As the used chamber was a collapsible bag, the volume of the chamber decreased over time due to the air taken by the instruments. [...]

45 **p.3.1.9.:** 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?

5 We were not able to separate, or at least qualitatively justify that the background factor, e.g. FE4
background, as this factor also included ions related to the car exhaust, thus being slightly mixed.
The lack of identified compounds for our gas phase measurements hinders the detailed analysis and
10 identification of possible background factors. In principle, by adding more factors, we might be able
to find one or more background factors, but as we would not have been able to identify those
explicitly, we did not apply this approach here but rather acknowledged the possible uncertainty in
our factors. Therefore, also quantifying this effect here is impossible.

15 **p.3.1.18.:** 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?

20 The number of observations which needed to be elevated to small positive values was not large, and
so was the number of time points. Removal of one negative value from one variable would require
the removal of the complete row, and as the negative values occur in a few variables, but not at the
same time point, this would remove a significant amount of the data. The small negative values in
the data are in cases when the concentration of the compound was in practice zero, but the
application of the baseline correction caused them to be very small negative numbers instead. We
acknowledge the possibility of increased uncertainty due to this kind of procedure, but due to the
25 very small size of the AMS data set, omitting the negative data points was not possible. PMF can be
actually applied to some extent even with the existence of some negative values, but the NMF
algorithm we used, does not allow the input of any negative values. In addition, also for PMF the
data should be positive by definition and without extensive inspection of the source code of the
PMF Evaluation tool we do not know how the software handles the negative numbers. Thus, to give
30 all SDRTs the same data as an input, we replaced the negative values. To acknowledge this issue
for the AMS data more clearly, we modified the end of the section 4.2.5 (please also note that this
section has now been modified to include the 2-factor solution from PMF instead of the 4-factor
one)

page 24 lines 17-23

35 [...] This, on the other hand, assigns different emphasis between the matrices, possibly causing
NMF to use more effort to reconstruct the data matrix with factor time series. However, the reader
should keep in mind that for detailed chemical analysis of such a small data set, especially with
PMF, the downweighing is advisable. In addition, the replacement of very small negative numbers
40 with very small positive numbers is not mandatory for PMF, as it can run containing a few negative
values. However, we did the replacement here, as the NMF algorithm used here, does require
strictly positive input data. Acquiring a balance between statistically good results and realistic
factors might be challenging, and to achieve more robust results, testing different error schemes
may be beneficial, especially for a data set with such a small size. [...]

45 **p.5.1.30:** “[...]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

50 In PCA, it is customary to scale columns of X so that in the scaled matrix, all of the
columns have the

5 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” [...]

10 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?

The sentence to which the reviewer is referring, explains how the scores (factor time series) are
usually calculated. However, in the end of the paragraph, we say “the scores are calculated without
standardizing the data matrix to achieve more interpretable component time series”.

15 However, we assume the paper from Paatero and Hopke means if the data matrix X is scaled before
applying PCA (or inside the algorithm). We used 2 variations of PCA, EVD-PCA, which uses
eigenvalue decomposition into the correlation matrix (calculated from unscaled X, though centering
for example, would not have an effect on the correlations), and SVD-PCA, which uses singular
20 value decomposition into the data matrix X or the correlation matrix. Thus, no scaling, whatsoever,
was applied when using EVD-PCA (which produced interpretable results shown in the manuscript).
We also tested SVD-PCA, and applied it to the scaled data matrix, as this was the recommendation
given for the algorithm we used. The SVD-PCA results were only shown in the supplement as they
were not interpretable. However, as the reviewer explains, this might affect negatively into the
25 results and is contradicting to the recommendations given by Paatero and Hopke (2003). Therefore,
we now also looked into the SVD-PCA applied to the unscaled data matrix X, but those results did
not show any improvement. We added one example of this to the supplementary material (see
section S4.1 **Figure S13**) and modified sections 4.1.2 and S4.1 accordingly. We modified sections
3.1.1 and 3.1.2 to emphasize that no scaling was applied to the data for EVD-PCA at any point and
30 explain scaling/not scaling situation for SVD-PCA.

section 3.1.1

page 5 lines 23-25

35 [...] It should be noted, however, that Eq. (1) describes the theory behind the PCA model, not the
actual calculation process, which is described below. Thus, for example, centering of variables is
not required. [...]

page 5 lines 29-31

40 [...] In our study we applied EVD-PCA to the correlation matrix (calculated from the unscaled data
matrix) and SVD-PCA was applied to the data matrix without and with the scaling (centered and
scaled by their standard deviations). In addition, the acquired [...]

section 3.1.2, page 7 lines 6-8

45 [...] As for the PCA explained in the previous section, Eq. (4) here describes the form of an EFA
model based on literature, not the direct calculation in the algorithm. Thus, no scaling (what e.g. the
subtraction of the mean \bar{y}_1 from \mathbf{y}_1 in Eq. (4) essentially is) is applied here. [...]

section 4.1.2, page 16 lines 10-14

50 [...] SVD-PCA (when applied to the scaled data matrix) was not able to separate α -pinene as its
own component, but instead created two factors which were dominated by the unreacted α -pinene
and its fragments (see Fig. S12 in SI section S4.1). In addition, the unrotated solution included a

large number of negative loadings, which complicated the interpretation of the components. No improvement was achieved when SVD-PCA was applied to the data matrix without any scaling (see. Fig. S13). [...]

5 section S4.1, page 6 lines 14-17
[...] components from SVD-PCA (applied to the scaled data matrix) and the unrotated component time series and original loadings (scaled eigenvalues) are shown in Fig. S12 with 4 components. The unrotated results from SVD-PCA when applied to the unscaled data matrix are shown in Fig. S13, and Fig. S14 shows the original loading values for the Oblimin rotated EVD-PCA with 4-
10 components. [...]

p.5.1.24. : 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)?

15 In PCA and EFA the rotation is applied to the loading matrix only, as the loading's time series are calculated after adjusting the actual loading values. This means in principle, that also the time series get more "different/separated" meaning rotation does not degrade, but in contrast even increase, the time series interpretability. In general, as pointed out, rotations potentially simplify the chemical
20 interpretation of the factors, but they might decrease the overall fit, as the rotation is applied after the solution has reached the optimal value for the condition to be minimized.

p.6.1.14.: 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).

25 Please see also our replies above ("p.5.1.30" and "p.5.1.24"). Regarding the error weighing. In EFA (or in PCA) no kind of weighing is applied based on measurement error as it is done in PMF. We now noticed from our formulation for EFA, that we use the term "the error term ε_1 ", which might cause misunderstanding. The term in question is not the error in the same sense as for PMF (which
30 needs the error matrix), but a residual error arising from the model fit. We renamed it to "the residual term ε_1 " to avoid further confusion.

Regarding the signal-to-noise, please see our reply to comment "p.7.1.21" in [page 9](#).

35 **p.6.1.6.:** 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?

40 After applying PCA to the correlation matrix, the acquired eigenvectors were scaled by the square root of the eigenvalues (see page 5, lines xx-xx) to acquire similar loading values compared to EFA. The limit of 0.3 given here, was applied to those loading values, which are (and should be in most cases) between -1 and 1. It does not mean that signals explaining less than 0.3 μm^3 are set to zero, but that the contribution of a variable into that specific factor is very small, and when setting the loading value to 0 for that factor, we enhance the separation between the factors. The limit of
45 0.3 for the absolute value of the loadings is widely applied in literature, and was thus selected (e.g. Field, 2013, Izquierdo et al., 2014). We adjusted the paragraph and references to justify our selection of the limit 0.3:

section 3.1.1, page 6 lines 26-29

[...] Very small loading values (absolute value less than 0.3) are suppressed to zero to enhance the separation of ions between the components. The limit of 0.3 was selected, as this is often given as a reference value for insignificant loadings (see. e.g. Field, 2013; Izquierdo et al., 2014). [...]

5 Regarding the mass signal conservation. As for EFA and PCA the factor scores can be calculated in various ways, and as the loading values can be negative, EFA and PCA do not conserve the original mass signal in that sense, but instead aim to conserve as much information (i.e. we could calculate how much of the correlations are preserved with different number of factors) from the original data as possible. Even though there are no practical obstacles to back-calculate the original signal from the obtained loadings and scores (which are calculated with the loadings), it does not make sense in a similar manner as for NMF and PMF, where the “scores” and “loadings” are optimized at the same time inside the algorithm. Due to this, the concentration of the factors between these methods is not fully comparable, but this is not an issue in this type of analysis.

15 **p.7.1.15** On the error model of $\text{st.dev} / \sqrt{n}$: 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)?

20 The PMF model was first introduced by using the standard deviations as an error metric, see Paatero and Tapper, (1994). The error metric applied here ($\text{st.dev} / \sqrt{n}$) is commonly called “standard error”. It takes into account that deductions from measurement with less observation contain more uncertainty. As the standard deviation may be high in this type of data (containing the step change at the onset of photochemistry), it may be a too conservative estimate for uncertainty in PMF analysis. Accounting for the number of observations gives a more reasonable estimate for total uncertainty. We clarified this in section 3.1.3:

30 page 8 lines 8-11

[...] But as a reference, the standard error of the mean (the standard deviations of the ion traces divided by the square root of the number of observations, i.e., length of the ion time series) was used as an error for both PTR-MS and AMS data. It considers that measurements with less observations contain more uncertainty. These error values are constant for each ion throughout the time series [...]

We did use the “standard” AMS error in our study (named Poisson-like error). For more details see our reply to comment (“p.7.1.26”) in [page 10](#).

40 **p.7.1.21.** (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):

45 *“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*

50

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.

10 We have added more justification for omitting the downweighing. Please also see our reply to Reviewer#1 comment ("page22 line 20")

section 3.1.3, page 8 line 27 – page 9 line 2

15 [...] . In contrast to the suggested best practice (Paatero and Hopke, 2003), we did not downweighed any ions in our data sets. This approach was used in order to give each SDRT an equal starting point for the analysis, as e.g. for NMF or PCA similar downweighing is not possible, as we do not have any error estimates to calculate the signal-to-noise ratios in similar manner. However, to avoid misguiding the reader to omit recommended data pre-processing practice for PMF, we also tested PMF with downweighing. This, as expected, did not change our results significantly, but we acknowledge it should be indeed applied if aiming for a more detailed chemical interpretation of the PMF factors. [...]

section S2.3, page 5, lines 1-5

25 [...] However, no removal or downweighing of the variables was applied in the results presented in the main manuscript, as for the weak ions the error values were already in the range of the ion signal itself, and for the bad variables the error was usually way above the signal throughout the ion time series. In addition, the number of weak or bad ions for both AMS and PTR-MS data were rather small. To justify our decision, we did run the PMF analysis with the downweighing. This, as expected, did not change our results or interpretation, and thus those results are omitted. [...]

30 Regarding the AMS error, see our reply to the next comment (to comment "p.7.1.26").

35 **p.7.1.26.**, 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.

40 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.

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

We did use the standard AMS error model, as mentioned in our reply to major comment in [page 3 line 27](#). We have now modified the manuscript accordingly to avoid this misunderstanding. In addition, we renamed the "Poisson style error" to "standard AMS error" to emphasize this.

Section 3.1.3, page 8 lines 17-25

[...] For AMS, we also applied a standard error that is frequently used by the AMS community. The standard AMS error consists of the minimum error related to the duty cycle of the instrument and the counting statistics following the Poisson distribution (Allan et al., 2003; Ulbrich et al., 2009).

5 Shortly, the standard AMS error for signal I can be formulated as

$$I_{err} = \alpha \sqrt{\frac{I_O + I_C}{t_s}}, \quad (9)$$

where α is an empirically determined constant (here $\alpha = 1.2$, generated by the AMS analysis software PIKA, <http://cires1.colorado.edu/jimenez-group/ToFAMSResources/ToFSoftware/index.html>, last visit on 9.9.2019), I_O and I_C are the raw signal of the ion of particle beam (ions s^{-1}) for the chopper at open and closed position, respectively, and t_s is the sampling time at a particular m/z channel (s). [...]

10

The signal following error was indeed tested for AMS data, but the results were omitted as they did not show any clear improvement and were rather similar to those acquired with the standard AMS error and thus would have lengthened the manuscript unnecessarily. We now mention this in the manuscript section 4.2.4:

15

page 23 lines 29-31

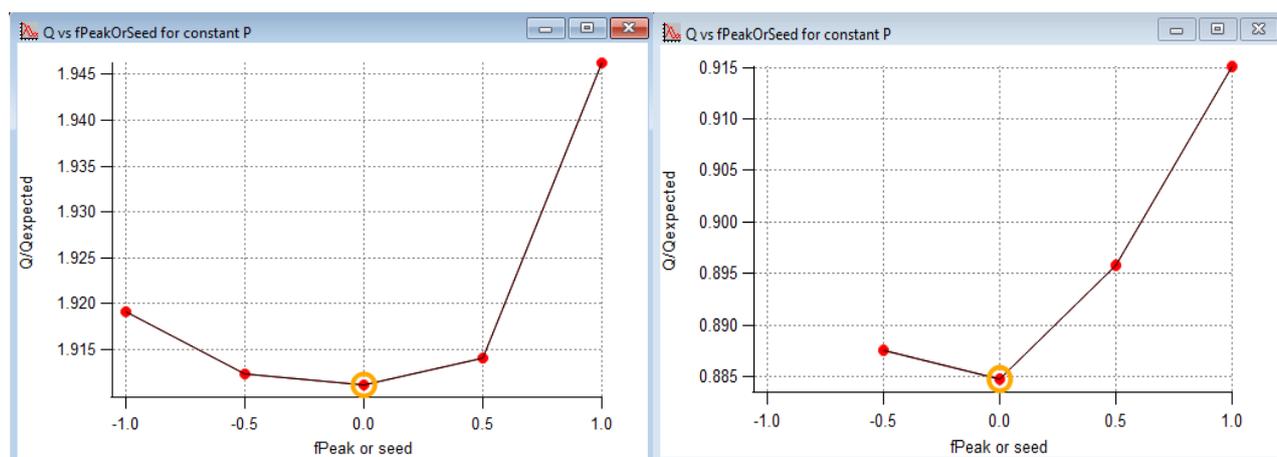
[...] The signal following error, used for PTR-MS, was also tested for particle phase data. However, as this type of error showed very similar behaviour as a time series as the Standard AMS error, and produced very similar outcome, those results are omitted from this manuscript. [...]

20

p.8.1.6.: How much does the Q (or Q/Q exp ratio) increase with these f_{peak} values?

Figure 1 below show the Q/Q_{exp} for PTR-MS data with the 5-factor solution with static error (left) and signal following error (right). The change with f_{peak} is rather small, as seen from the figures. The actual factorization also does not change (for PTR-MS) with different f_{peak} , indicating the 5-factor solution is quite robust.

25



30

Figure 1 Q/Q_{exp} for PTR-MS data with the 5-factor solution with static error (left) and signal following error (right).

p.9.1.10.: How do you interpret the negative loadings in EFA and PCA? Why were these solutions deemed physically sound, if they feature negative mass loadings?

5 These negative loadings refer to the feature of EFA and PCA that the factor loadings can be negative. Negative factor loading may have different interpretations. It may indicate that the compound has a decreasing effect on the factor, i.e. it acts as a sink for the compounds with positive loading in the same factor. In chamber experiments, negative loading may also refer to decreasing concentration of the compound within chemical reactions. For example, a compound that can be rather easily oxidized, or forms in the later state of the experiment, might get negative loading in a factor, that mostly contains later generation products. An example of this is benzene (detected as $C_6H_7^+$), assigned to FE1. When inspecting the original loading values, it has a negative loading in FE1 (identified as later/slowly forming products), and a positive loading in FE4 (identified as precursors from car exhaust/background). As benzene originates from the car exhaust, it contributes positively to FE4. However, as it oxidizes over the course of the experiment (thus it has a decreasing concentration), it contributes negatively to FE1, which mostly includes those later generation products.

15 We understand our approach does indeed conceal some additional information EFA/PCA can provide, but as we did not have enough detailed information of the measured compounds (e.g. the sum formulas were not identified for all compounds), the interpretation of the negative loadings separately, for the most part, was not possible. We added more clarification about this in the end of section 3.1.5, and also discuss this property of EFA/PCA in the [new section 4.4](#) (see [pages 1-3 in this reply](#)).

section 3.1.5, page 10 lines 9-19
25 [...] However, this type of approach conceals the information of the negative factor loadings in EFA and PCA (which are included in the calculation of factor/component time series as weights), but instead visualizes the general contribution of an ion to a factor. Negative factor loadings may have different interpretations. It may indicate that the compound has decreasing effect on the factor, i.e. it acts as a sink for the compounds with positive loading in the same factor. In chamber experiments, negative loading may also refer to decreasing concentration of the compound participating in chemical reactions if it acts as a precursor for other compounds in the same factor. One example of this is benzene (C_6H_6), observed with PTR-MS. When inspecting the original loading values from EFA, for example, it has negative loading in FE1 (identified as later/slowly forming products), and positive loading in FE4 (identified as precursors from car exhaust/background). As benzene originates from the car exhaust, it contributes positively to FE4. However, as it oxidizes over the course of the experiment (thus has decreasing concentration), it has a strong correlation with oxidation products but appears negative in FE1, which mostly includes those later generation products. [...]

40 **p.10.1.5.:** “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.

45 Obvious disadvantage is in cases where one compound affects multiple factors by physical or chemical definition. Then clustering loses information by forcing the compounds on only one cluster. For example, when identifying pollution sources, black carbon can be emitted from both wood burning emissions and traffic. There, clustering might not be the best approach. On the other hand, computational time (as shown also in this study) is one clear advantage for PAM or hard

division techniques in general, especially when analyzing very long ambient data sets. Hard division techniques have also been shown to work efficiently when distinguishing between different coffee types (i.e. strict separation between clusters is needed), as shown by hierarchical clustering by Sánchez-López et al. (2014).

We added few examples to section 3.2.1, page 11 lines 15-21

[...] One obvious advantage of cluster analysis (or hard division techniques in general) is computational time, especially if analyzing long ambient data sets. For laboratory measurements, this most likely is not an issue. Hard division techniques have also been shown to work efficiently for VOC measurements when distinguishing between different coffee types (espresso capsules), where strict separation between clusters is needed, as shown in Sánchez-López et al., (2014). For source apportionments studies, where one variable might originate from multiple sources, cluster analysis using hard division technique is probably not as suitable as softer division techniques, which can assign one variable to multiple sources/factors. [...]

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 and f2 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.

Following the suggestion regarding the PMF solution for AMS in comments below (see reply to comment “p.21., Section 4.2.4” in page 15) we now present a 2-factor solution in the PMF results. The figure below shows the evolution of the two PMF factors and α -pinene. In the chamber, we assume a steady-state OH concentration of 10^7 molecules/cm³, and the reaction rate constant OH- α -pinene of 52.3×10^{-12} cm³ molecule⁻¹ s⁻¹, it is very likely to get α -pinene consumed (25ppb) in a time scale of half an hour. If we assume a mass yield of α -pinene SOA of 30% in this study (Kari et al., 2019), the consumed α -pinene (about 140 $\mu\text{g m}^{-3}$) can be transformed to a SOA mass concentration of 52 $\mu\text{g m}^{-3}$. This is consistent with what is observed in the figure. Thus, we can assume the time series from the SDRTs used in this study reflect the reaction kinetics in the chamber and the processes take place on reasonable timescales.

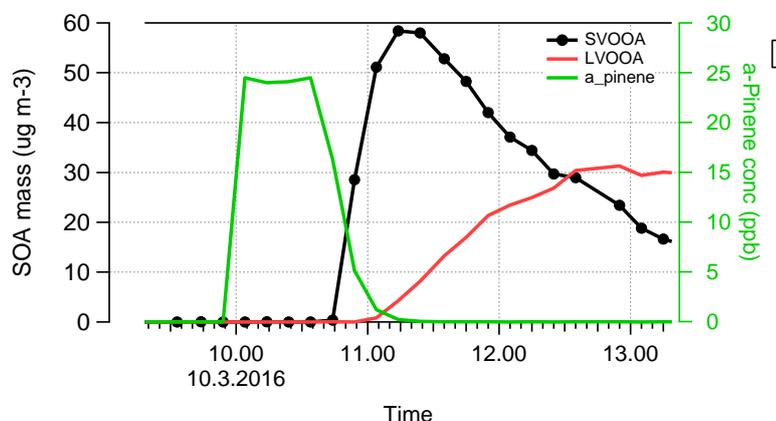


Figure 2 Two-factor solution for particle-phase data from PMF (fpeak = 0) with the injected α -pinene.

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.

Figures 1, 2 and 11 are now modified as suggested, RSS in Figure 2 and 11 is changed to more NMF related metric as suggested by Reviewer #1. Captions in each figure were adjusted to include the information of the data (AMS or PTR-MS), and we added factor description (those in table 1 & 2) to the figures in the manuscript (when applicable), as requested by the reviewer#1.

With the factor contribution figures, we looked for a way of displaying the factor/compound “compositions” in a way allowing direct comparison of the different methods. Normalizing the factor contribution for each ion was chosen to avoid the very different weighting methods in NMF and PMF between the factorization matrices and also to be able to show EFA/PCA/PAM results in a similar manner. We now added figures with the factor mass spectra in separate panels in the SI material for easier comparison (**section S5, renamed to “Contrast angle and factor spectra comparison”, Figures S20, S21**), and pointed them out in the manuscript:

section 4.1.6

page xx, lines xx-xx

[...] Table 1 summarizes the acquired results from different SDRTs for the gas phase composition data measured with PTR-MS, and Figs. S20 and S21 in SI section S5 show separate factor contributions for each of the SDRTs. [...]

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.

Please see our reply to specific comment (“p.9.1.10”) in **page 12**. In addition, see our reply to the major comment in the beginning of this reply (we have now included a **new section 4.4**, to the manuscript with more discussion of the objectives or the analyses).

p.17.1.25. “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.7.1.15. 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.

We did use the “standard” AMS error in our study (named Poisson-like error). For more details see our reply to comment (“p.7.1.26”) in **page 10**.

In principle the feature of having larger errors for fast changes is intentional, as if we would have ambient data, such large variations most likely would be considered as instrument malfunctions (“spikes” in the data), thus applying a larger error for those data points is reasonable. Our aim using the “signal following error” was indeed to use an error scheme, that could also be used for ambient measurements. In addition, for ambient data we are not that interested in the brief period with rapid and fast changes, as long as the new values after these changes are captured well by PMF. However, in chamber measurements this type of changes can be relatively large (as in this study) and should be assessed in a different way. We wanted to test two very different error schemes here to investigate how this affects to the results. The static error stays constant over time, thus giving equal weight to time points whereas the signal following error assumes those fast changes are not so important. For example, if we would be interested in long term changes in ambient measurements, signal following error would be the choice. We have now reworded the corresponding part in section 4.1.5 to be consistent:

page 19, lines 17-24

[...] This is caused by the used error scheme, where errors are larger for the fast changes in the data (Fig. S4b). In ambient data not measured at instant proximity of strong emission sources, for which PMF is often used, this type of error is beneficial as there the fast changes are more likely to be noise or instrument malfunctions (excluding, for example, sudden primary emission plumes), and we are more interested of the long-term changes instead. For laboratory data, where large changes are often caused by rapid changes in actual experimental conditions, e.g. due to injecting α -pinene or turning the UV lights on, the static type of error is most likely preferable. Usage of the static error scheme helps to avoid overcorrecting intentional (large) changes in experimental conditions and confusing them with real variation taking place during the experiment and typically being much less pronounced. [...]

p.21., Section 4.2.4.: I would have inspected the 2 factor solution, is it had percentage-wise the largest decrease in Q/Q_{exp} . 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).

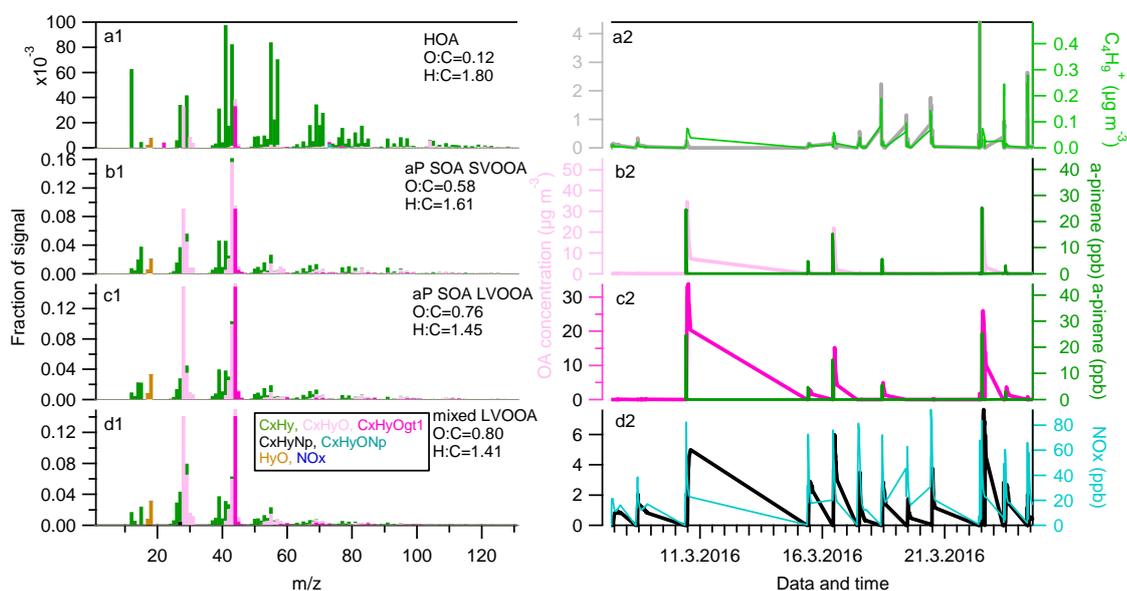
The analysis on the error scheme issues seems correct.

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 f_1, f_2, f_3) 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, $LV(f_1, f_2, f_3)/SV(f_4)$ split in 13 c&d.

Regarding EFA and PCA. We would not trust the number of factors – tests here, as the data from particle phase measurement has more variables than time points, which creates significant calculational issues (see e.g. page 21 lines 19-24) for the eigenvalue decomposition used with PCA and for the EFA algorithms due to the issues with the correlation matrix. We can still run EFA and PCA, but in both cases some smoothing is applied inside the algorithms due to the non-positive-definite correlation matrix. Solutions with 2 factors/clusters were already shown in Figs. S24 and S26/S27. The PCA solution with 2 components was omitted, as it was almost identical to the EFA solution thus bringing no new information to the manuscript.

We do agree with the reviewer about the interpretability of the presented PMF solution. We have conducted PMF on the entire experiment dataset (compare to only one day experiment in this

study), which produced four meaningful factors (see Figure 3 below). The 4-factor solution is nicely consistent with those in NMF (Sec. 4.2.3). HOA (hydrocarbon OA) was dominated by the prominent C_nHC_{2n-1} and C_nHC_{2n+1} family ions and appeared as primary emission in the experiments that included the vehicle, the α P_SOA_SVOOA and α P_SOA_LVOOA factors are present only in the pure α -pinene SOA experiments. The mixed_LVOOA factor appears in both mixed car exhaust and α -pinene SOA experiments. The four factors are physically and chemically meaningful. However, in this study, measurement data of only one day (i.e. one chamber experiment) was included for the analysis. This clearly caused issues for PMF, and as suggested by the reviewer, a two-factor solution may be preferable for this subset of data. Hence, we have now changed the AMS PMF results in the main manuscript to the 2-factor solution and added the results with 4 factors to the SI. The factors from 2-factor solution can be interpreted as SVOOA and LVOOA.



15 Figure 3 The mass profiles and time series of a 4-factor solution of PMF analysis on the entire experiment (Kari et al., 2019)

The results from PMF with particle phase data were modified, and the sections 4.2.5, 5 and the corresponding SI material, abstract and figures were modified accordingly.

20 corrected section 4.2.4

The Q/Q_{exp} for the two error schemes are shown in Fig. 11b. Neither of the error schemes show a clear inflection point. Examples of behaviour of the errors as a time series are shown in Fig. S7 (SI Sect. S2.2). With the standard AMS error, the Q/Q_{exp} values do not reach 1 (with 10 factors $Q/Q_{exp} = 1.76$) whereas with the static error the values decrease below 1 for 7 factors. The solutions with 2-5 factors were inspected, and the 2-factor solution (Fig. 13) is presented here as the most interpretable one (summarized in Table 2). The primary OA factor, separated by NMF (Fig. 12, FN2) was only found if using 4 factors and the static error scheme in PMF (see SI Sect. S6.2, Fig. S31a-b). However, interpretation of the time series for that solution was found to be very difficult due to the extreme anticorrelation between the time series, and thus the 2-factor solution was selected. The two factors were interpreted as SVOOA and LVOOA. In addition, the largest relative decrease in the Q/Q_{exp} was observed with the 2-factor solution.

The residuals for the standard AMS error were smaller, as shown in Fig. 14. This agrees with the analysis of the gas phase data set, where the residual for the signal following error (which has a similar profile in time as the standard AMS error) were generally smaller compared to the static error.

- 5 The signal following error, used for PTR-MS, was also tested for particle phase data. However, as this type of error showed very similar behaviour as a time series as the standard AMS error, and produced very similar outcome, those results are omitted from this manuscript.

corrected section 4.2.5

- 10 Table 2 summarizes the acquired results from NMF and PMF for the particle phase composition data measured with AMS. PAM was not able to separate distinct clusters due to the inability of clustering techniques to classify an ion into multiple clusters. Comparing the relative factor spectra and fraction of signal from NMF and PMF with the static error (Figs. 12b and 13b), the distribution of ions is similar between the LVOOA in the PMF solution and the mixed LVOOA factor in the NMF solution, and also similar between the SVOOA in the PMF solution and the α P-SOA factor (integrated- α P-SOA-SVOOA and α P-SOA-LVOOA factors). When inspecting the individual ion time series in the original AMS data, most of them have rather “smooth” behaviour, similar to the factors acquired from NMF. It seems, that PMF gives more weight to the background ions (with very small concentration) which do not have that clear structure in their time series, thus including more of their behaviour in the final factors, if the number of factors in PMF is increased from 2 (see SI Sect. S6.2 Figs. S30-S32). Residuals from NMF reconstruction (with 4 factors) were over ten orders of magnitudes smaller (for NMF between $-1.3 \cdot 10^{-13}$ and $7.1 \cdot 10^{-14}$) than those from PMF (between -0.06 and 0.13 for the 2-factor solution with standard AMS error, see Fig. 14), indicating better reconstruction of the data with NMF. Most likely PMF struggles with the small data set, thus not being able to recover all the factors found by NMF and construct reasonable time series for those factors (see 4-factor solution in Fig. S31), whereas NMF does not seem to be affected by the data size. In addition, the weighing between the factorization matrices between NMF and PMF is different not only due to the error matrix that is given as a weight in PMF, but also because of the different solving algorithms for each method. This, on the other hand, assigns different emphasis between the matrices, possibly causing NMF to use more effort to reconstruct the data matrix with factor time series. However, the reader should keep in mind that for detailed chemical analysis of such data set, especially with PMF, downweighing is advisable. In addition, the replacement of very small negative numbers with very small positive numbers is not mandatory for PMF, as it can run with a few negative values into some extent. However, we did the replacement here, as the NMF algorithm used here, does require strictly positive input data. Acquiring a balance between statistically good results and realistic factors might be challenging, and to achieve more robust results, testing different error schemes may be beneficial, especially for a data set with such small size.

abstract page 1 lines 23-24

- 40 [...] For the particle phase measurements, four factors were discovered with NMF: one primary factor, a mixed LVOOA factor, and two α -pinene SOA derived factors. PMF was able to separate two factors: SVOOA and LVOOA. [...]

section 5

- 45 page 26 line 32 – page 27 line 1

[...] From the particle phase data, NMF was able to separate four factors, whereas PMF separated two. [...]

page 27 lines 3-5

[...] EFA and PCA had computational constraints due to the small data size acquired from the AMS and could not to be applied. In addition, PMF also faced assumedly computational issues with the small particle phase data set, thus not being able to reasonably separate the HOA factor. [...]

Technical corrections:

p.3.1.4.: PTR generally refers to the ionization reaction and not the instrument, Please use PTR-MS or similar, common acronym.

We corrected PTR to PTR-MS in the manuscript and supplement and in all figures.

p.7.1.30. ts 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.

We have now clarified this in the manuscript. Please see our reply to specific comment (“p.7.1.26”) in [page 10](#).

p.21.1.13.: 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.

We added the reference that was missing (Zhang et al., 2005), corrected one other reference (Hoyle et al, 2011 to National Research Council, 1991) and corrected the numbering of the figures and indicate in each figure caption the type of data presented.

section 4.2.3, page 23 line 5

[...] contains marker ions associated with HOA (e.g. m/z 57, Zhang et al., 2005) condensed again in a cooler chamber [...]

REFERENCES

Field, A.: Discovering Statistics using SPSS, 4 ed., SAGE, London, 2013.

Izquierdo, I., Olea, J., and Abad, F. J.: Exploratory factor analysis in validation studies: uses and recommendations, *Psicothema*, 26, 395-400, 10.7334/psicothema2013.349, 2014.

Sánchez-López, J. A., Zimmermann, R., and Yeretjian, C.: Insight into the time-resolved extraction of aroma compounds during espresso coffee preparation: online monitoring by PTR-ToF-MS., *Anal Chem*, 86, 11696-11704, 2014.

Zhang, Q., Alfarra, M. R., Worsnop, D. R., Allan, J. D., Coe, H., Canagaratna, M. R., and Jimenez, J. L.: Deconvolution and quantification of hydrocarbon-like and oxygenated organic aerosols based on aerosol mass spectrometry, *Environ Sci Technol*, 39, 4938-4952, 10.1021/es048568l, 2005.

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.

5 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](http://dx.doi.org/10.1016/S0003-2670(02)01643-4)

10 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.

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

Sini Isokääntä¹, Eetu Kari^{1,a}, Angela Buchholz¹, Liqing Hao¹, Siegfried Schobesberger¹, Annele Virtanen¹, Santtu Mikkonen^{1,2}

5 ¹ Department of Applied Physics, University of Eastern Finland, Kuopio, 70210, Finland

² [Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, 70210, Finland](#)

^a [currently at: Neste Oyj, Espoo, Finland](#)

Correspondence to: S. Isokääntä (sini.isokaanta@uef.fi)

10 **Abstract.** Online analysis with mass spectrometers produces complex data sets, consisting of mass spectra with a large number
of chemical compounds (ions). Statistical dimension reduction techniques (SDRTs) are able to condense complex data sets
into a more compact form while preserving the information included in the original observations. The general principle of
these techniques is to investigate the underlying dependencies of the measured variables, by combining variables with similar
characteristics to distinct groups, called factors or components. Currently, positive matrix factorization (PMF) is the most
15 commonly exploited SDRT across a range of atmospheric studies, in particular for source apportionment. In this study, we
used 5 different SDRTs in analysing mass spectral data from complex gas- and particle phase measurements during laboratory
experiment investigating the interactions of gasoline car exhaust and α -pinene. Specifically, we used four factor analysis
techniques: principal component analysis (PCA), positive matrix factorization (PMF), exploratory factor analysis (EFA), and
non-negative matrix factorization (NMF), as well as one clustering technique, partitioning around medoids (PAM).
20 All SDRTs were able to resolve 4-5 factors from the gas phase measurements, including an α -pinene precursor factor, 2-3
oxidation product factors and a background/car exhaust precursor factor. NMF and PMF provided an additional oxidation
product factor, which was not found by other SDRTs. The results from EFA and PCA were similar after applying oblique
rotations. For the particle phase measurements, four factors were discovered with NMF ~~and PMF~~: one primary factor, a mixed
LVOOA factor, and two α -pinene SOA derived factors. [PMF was able to separate two factors: SVOOA and LVOOA.](#) PAM
25 was not able to resolve interpretable clusters due to general limitations of clustering methods, as the high degree of
fragmentation taking place in the AMS causes different compounds formed at different stages in the experiment to be detected
at the same variable. However, when preliminary analysis is needed, or isomers and mixed sources are not expected, cluster
analysis may be a useful tool as the results are simpler and thus easier to interpret. In the factor analysis techniques, any single
ion generally contributes to multiple factors, although EFA and PCA try to minimize this spread.
30 Our analysis shows that different SDRTs put emphasis on different parts of the data, and with only one technique some
interesting data properties may still stay undiscovered. Thus, validation of the acquired results either by comparing between

different SDRTs or applying one technique multiple times (e.g. by resampling the data or giving different starting values for iterative algorithms) is important as it may protect the user from dismissing unexpected results as “unphysical”.

1 Introduction

Online measurements with mass spectrometers produce highly complex data comprised of hundreds of detected ions. High resolution mass spectrometer enables identification of the elemental composition of these ions revealing chemical composition information about the sample. However, even with the highest resolution, mass spectrometers are not able to resolve isomers. Instead the large number of identified ions can make data interpretation challenging due to the sheer number of variables. Different statistical dimension reduction techniques (SDRT) were developed to compress the information from complex composition data into a small number of factors, which can be further interpreted by their physical or chemical properties. In other words, these methods are used to understand the underlying relationships of the measured variables (i.e., detected ions). Principal component analysis (PCA), which was introduced already in the beginning of the 20th century by Karl Pearson is probably the first SDRT, even if the modern formulation of PCA was introduced decades later (Pearson., 1901; Hotelling, 1933). In atmospheric studies, the most exploited method, especially in the analysis of long time series of aerosol mass spectrometer (AMS) data, is positive matrix factorization (PMF) developed in the mid-1990s (Paatero and Tapper, 1994). Other SDRTs that are widely applied in different fields of science for the analysis of multivariate data include PCA and exploratory factor analysis (EFA), which are popular especially in medical and psychological studies (Raskin and Terry, 1988; Fabrigar et al., 1999). In atmospheric studies, the latter methods have not gained as wide popularity, but a few examples still exist. Customized PCA was applied to organic aerosol data collected from Pittsburgh in 2002 (Zhang et al., 2005) and a more traditional version of PCA was used to analyze [ze](#) Chemical Ionization Reaction Time-of-Flight Mass Spectrometer (CIR-ToF-MS) and compact Time-of-Flight Aerosol Mass Spectrometer (cToF-AMS) data acquired in smog chamber studies during several measurement campaigns (Wyche et al., 2015). Additionally, EFA and PCA have been applied in several source apportionment studied in the environmental science fields (Pekey et al., 2005; Sofowote et al., 2008), and a recent study on plant VOC emissions applied EFA to separate effects of herbivory induced stress from the natural diurnal cycle of the plants (Kari et al., 2019a). Very much [a](#)like PMF, non-negative matrix factorization (NMF) is one of the most used methods in the analysis of DNA microarrays and metagenes in computational biology (Brunet et al., 2004; Devarajan, 2008), but NMF has also been applied in atmospheric studies (Chen et al., 2013; Malley et al., 2014). Comparisons between the performance of some of the SDRTs presented in this paper already exists, but due to the popularity of PMF, other methods are not applied as widely in atmospheric studies. As EFA and PCA are rather similar methods, and they have also existed many decades, multiple comparisons between them exists, especially in the medical and psychological research fields (see. e.g. Kim, 2008). The introduction of PMF has also inspired comparison studies between PMF and EFA (Huang et al., 1999), and PMF and PCA were already shortly compared upon publication of PMF as the positivity constraints were presented as an advantage over PCA (Paatero and Tapper, 1994). Although PMF has been shown to be a very powerful

tool in the analysis of environmental AMS data from field studies (e.g. Ulbrich et al., 2009; Zhang et al., 2011; Hao et al., 2014, Chakraborty et al., 2015), it has not been applied as widely in laboratory and smog chamber research (Corbin et al., 2015; Kortelainen et al., 2015; Tiitta et al., 2016; Koss et al., 2020). Latest studies have applied PMF also to chemical ionization mass spectrometry data (Yan et al., 2016; Massoli et al., 2018; Koss et al., 2020) which is able to resolve more oxidized compounds. The special conditions in lab experiments (sharp change at the beginning of experiments, e.g., switching on UV-lights) ~~may present an additional test scenario challenges~~, as PMF ~~has been mostly used~~ ~~was originally developed~~ for field measurement data sets where ~~the main focus is often in the long-term trends and~~ real changes in factors are expected to be ~~much slower~~ ~~more subtle~~ than e.g. ~~the variations in~~ the noise in the data. In addition, field measurements commonly yield very large data sets, including thousands of time points, whereas laboratory experiments may be much shorter. Recently, scientists from atmospheric studies have been motivated to test and adapt other techniques and algorithms to reduce the dimensionalities of their data, in addition to the more “traditional” version PMF introduced in the 90s. For example, Rosati et al., 2019 introduced correlation based technique for multivariate curve analysis (similar to NMF) in their analysis of α -pinene ozonolysis. Cluster analysis has been applied in a few studies. Wyche et al. (2015) applied hierarchical cluster analysis (HCA) to investigate the relationships between terpene and mesocosm systems. In the study from Äijälä et al. (2017), they combined PMF and k-means clustering to classify and extract the characteristics of organic components. In addition, very recent paper by Koss et al. (2020) also compared the dimension reduction abilities of HCA and gamma kinetics parametrization to PMF when studying mass spectrometric data sets.

In our study we chose a set of SDRTs having fundamental differences. For example, PMF usually splits one ion into several factors, whereas most of clustering techniques assign one ion into one cluster only. If isomers with the same chemical composition but different functionality are expected, splitting of ions into several factors might be preferred. On the other hand, clustering might be more suitable for a more simplified or preliminary approach ~~(as it is computationally less demanding)~~, or when the chemical compounds in the data are already known ~~or if strict division between variables is preferred~~. In this study, we validate the usability of the chosen SDRTs in laboratory studies for two different mass spectrometer devices, PTR-ToF MS and AMS (gas- and particle phase composition), and different data sizes due to different measuring periods and time resolutions. ~~Further, we examine the performance of the SDTRs when the data includes large and rapid changes in the composition.~~

2 Experimental data

The datasets investigated in this study ~~were~~ gathered during experiments conducted as part of the TRACA campaign at the University of Eastern Finland. A detailed description of the experimental setup and reaction conditions can be found in Kari et al. (2019b). Briefly, the measurement setup consisted of a modern gasoline car (VW Golf, 1.2 TSI, Euro 6 classification) which was driven at a constant load of 80 kilometres per hour after a warm up period with its front tiers in a dynamometer. The exhaust was diluted using a two-stage dilution system and fed into a 29 m³ ~~collapsible~~ environmental PTFE chamber

ILMARI (Leskinen et al., 2015). For the experiment investigated in this study, α -pinene ($\sim 1 \mu\text{L}$, corresponding to 5 ppbV) was ~~injected~~ added in to the chamber to resemble biogenic VOCs in typical suburban area in Finland. Atmospherically relevant conditions were simulated by adding O_3 to convert extra NO from vehicle emissions to NO_2 and adding more NO_2 to the chamber if needed. With these additions, atmospherically relevant VOC-to- NO_x ($\sim 7.4 \text{ ppbC/ppb}$) and NO_2 -to-NO ratios were
5 achieved to resemble the typical observed level in suburban areas (National Research Council, 1991). Hoyle et al., 2011 by adjusting the VOC to NO_x ratio and adding O_3 to adjust the NO_2 to NO ratio (see Table S1) to resemble the typical observed level in urbanized areas (Hoyle et al., 2011). Chamber temperature was held constant at $\sim 20^\circ\text{C}$ and relative humidity was adjusted to $\sim 50\%$ before the start of the experiment. Blacklight (UV-A) lamps with a light spectrum centred at 340 nm were used to form OH radicals from the photolysis of H_2O_2 . The start of photooxidation by turning on the lamps is defined as
10 experiment time 0 in the following. Vertical dashed lines in the figures indicate α -pinene injection and the start of photo-oxidation, respectively. A short summary of the Eexperimental conditions and the behaviour of the injected α -pinene as a time series are shown in the SI material (Sect. S1)

Volatile organic compounds (VOCs) in the gas phase were monitored with a proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF MS 8000, Ionicon Analytik, Austria, hereafter referred to as PTR-MS). Typical concentration for few
15 example VOCs in the mid-way of the experiment were $2 \mu\text{m}^3$ for toluene, $0.2 \mu\text{m}^3$ for TMB (trimethylbenzene) and $1.7 \mu\text{m}^3$ for $\text{C}_4\text{H}_4\text{O}_3$. Detailed setup, calibration procedure and data analysis of the used high resolution PTR-MS have been explicitly presented in Kari et al. (2019b). In the campaign, the high mass resolution of the instrument (>5000) enabled the determination of the elemental compositions of measured VOCs. The instrumental setting intended to minimize the fragmentation of some compounds, so the quantitation of the VOCs was possible. The chemical composition of the particle
20 phase of the formed SOA was monitored with a soot particle aerosol mass spectrometer (SP-AMS, Aerodyne Research Inc., USA, hereafter referred to only as AMS, Onasch et al., 2012). In brief, the SP-AMS was operated at 5 min saving cycles, alternatively switching between the electron ionization (EI) mode and SP mode. In EI mode, the V-mode mass spectra were processed to determine the aerosol mass concentration and size distribution. The mass resolution in the mode reaches ~ 2000 . The SP-mode mass spectra were used to obtain black carbon concentration. As the used chamber was a collapsible bag, the
25 volume of the chamber decreased over time due to the air taken by the instruments. For the experiment investigated in this study, both gas- and particle phase data were analyzed with all SDRTs (Sect. 4.1 and 4.2). However, due to the small data size for the particle phase, not all SDRTs were applicable.

In contrast to the ~~PTR~~PTR-MS data used in Kari et al. (2019b), we did not apply baseline correction to the data. Overestimation of the baseline correction may cause some of the ions with low signal intensity to have negative “concentration”, which is not
30 physically interpretable. Also, negative data values cause problems for some SDRTs, as e.g. PMF and NMF need a positive input data matrix. In addition, SDRTs should be able to separate background ions into their own factor, meaning it is not mandatory to remove them before applying SDRTs. This approach will cause some bias to the absolute concentrations of the ions and resulting factors, but as we are more interested in the general division of the ions to different factors, and their behaviour as a time series when comparing the SDRTs, it does not significantly affect our interpretation of the results. All

recommended corrections (including baseline subtraction, Ulbrich et al., 2009) were applied to the AMS data. As the processed AMS data is always the difference between the measured signal with and without particles, negative values are possible if the particle free background was elevated. In the investigated dataset, only few data points exhibited slightly negative values. Thus, it was possible to set these data points to a very small positive value ($1 \cdot 10^{-9}$) to enable the analysis with SDRT methods without a significant positive bias in the data. [In addition, as the main focus of our study was to compare the performance of the different SDRTs with different types of mass spectra, instead of detailed analysis of the chamber experiment, we have also included the pre-mixing period during the \$\alpha\$ -pinene \(i.e. \$t < 0\$ \) injection into our analysis.](#)

3 Dimension reduction techniques

3.1 Factorization techniques

10 3.1.1 Principal Component Analysis (PCA)

PCA is a statistical procedure where the variables are transformed into a new coordinate system. The first principal component accounts for the most variance of the observed data and each succeeding component has then the largest possible variance with the limitation that the component must preserve orthogonality to the preceding component. In other words, PCA seeks correlated variables and attempts to combine them into a set of uncorrelated variables, i.e., principal components, which include as much as possible of the information that was present in the original observations (Wold et al., 1987; Morrison, 2005; Rencher and Christensen, 2012; Tabachnick and Fidell, 2014). The principal components ~~can be~~ [are often](#) described by a group of linear equations, where e.g. the first principal component \mathbf{c}_1 (table of the used mathematical symbols and notations is presented in the appendix) can be presented as

$$\mathbf{c}_1 = a_{11}\mathbf{y}_1 + a_{12}\mathbf{y}_2 + \dots + a_{1m}\mathbf{y}_m, \quad (1)$$

20 where the coefficients a_{1j} ($j = 1, \dots, m$) are normalized characteristic vector elements assigned to the specific characteristic root of the correlation matrix \mathbf{S} , and \mathbf{y}_i ($i = 1, \dots, m$) are the centred variables (Morrison, 2005; Rencher and Christensen, 2012). As the responses in the first principal component have the largest sample variance, $s_{y_1}^2$, for all normalized coefficient vectors applies

$$s_{y_1}^2 = \sum_{i=1}^m \sum_{j=1}^m a_{1j} a_{1j} s_{ij} = \mathbf{a}_1^T \mathbf{S} \mathbf{a}_1, \quad (2)$$

25 where $\mathbf{a}_1^T \mathbf{a}_1 = 1$ (Morrison, 2005). The number of principal components is equal to the number of variables (m) in the data minus one, and p components are selected to interpret the data. [It should be noted, however, that Eq. \(1\) describes the theory behind the PCA model, not the actual calculation process, which is described below. Thus, for example, centring of variables is not required.](#) To find the principal components, either eigenvalue decomposition (EVD) or singular value decomposition (SVD) can be used. Mathematical formulation of EVD and SVD can be found from Golub and Van Loan, 1996). EVD is applied to the correlation or covariance matrix \mathbf{S} whereas SVD can be applied also to the observed data matrix directly. Often, 30 due to this difference, SVD is considered as its own method instead of describing it as a variation of PCA. Here, however, it

is referred to as SVD-PCA. In our study [we applied EVD-PCA to the correlation matrix \(calculated from unscaled data matrix\) and SVD-PCA was applied to the data matrix without and with the scaling \(centred and scaled by their standard deviations\).](#) [In addition,](#) the acquired eigenvectors and vectors corresponding to the singular values were scaled by the square root of the eigenvalues/singular values to produce loading values (i.e., contribution of a variable to a component) more similar to those obtained in exploratory factor analysis (EFA). The PCA analysis was performed in R statistical software with the addition of the psych-package (Revelle, 2018; R Core Team, 2019).

The acquired principal components can be rotated to enhance the interpretability of the components. Rotations can be performed in orthogonal or oblique manner, where the orthogonal methods preserve the orthogonality of the components, but oblique methods allow some correlation. However, rotation of the principal components does not produce another set of principal components but merely components. By original definition (Hotelling, 1933), only presenting unrotated solution is considered as principal component analysis, but later formulations allow also orthogonal rotations (Wold et al., 1987). Though there are no computational restrictions for applying oblique rotation on components, the restriction is only definitional, as the original principal components were presented as orthogonal transformation. In any case, rotated solutions do not fulfil the assumption of principal ordering of components. In this study, orthogonal Varimax rotation, which maximises the squared correlations between the variables, and oblique Oblimin rotation were used to increase large loading values and suppress the small ones to simplify the interpretation (Kaiser, 1958; Harman, 1976).

Multiple ways exist to calculate the PCA component scores (i.e., component time series). In general, the components scores are calculated as

$$\mathbf{F} = \mathbf{XB} \tag{3}$$

where \mathbf{X} includes the analysed variables ([often](#) centred and scaled by their standard deviations), and \mathbf{B} is the component coefficient weight matrix (Comrey, 1973). One simple way to calculate the component scores is to use the component loading values directly as weights. This approach is often referred to as a sum score method. Depending on the application, the loadings can be used as they are, they can be dichotomized (1 for loaded and 0 for not loaded), or they can be used as they are but suppressing the low values by some threshold limit. We applied the latter method here, as the dichotomized loadings (i.e., one ion stems only from one source or source process) seldom describe true physical conditions in nature. In a case when the data items are on the same ~~scale~~ unit, the data may be used without standardization (Comrey, 1973). As this is the case in all of our respective data sets (concentration units for [PTRPTR-MS](#) data are ppb and $\mu\text{g}/\text{m}^3$ for AMS), the scores are calculated without standardising the data matrix to achieve more interpretable component time series. [This applies for both EVD-PCA and SVD-PCA component scores.](#) Very small loading values (absolute value less than 0.3) are suppressed to zero to enhance the separation of ions between the components. [The limit of 0.3 was selected, as this is often given as a reference value for insignificant loadings \(see. e.g. Field, 2013; Izquierdo et al., 2014\).](#) The components from PCA in the results sections are labelled with CO.

3.1.2 Exploratory Factor Analysis (EFA)

Similar to the EVD-PCA which takes an advantage of the correlations between the original variables, EFA generates the factors trying to explain the correlation between the measured variables (Rencher and Christensen, 2012). For a data matrix \mathbf{X} having m variables (ions) and n observations (time points), the EFA model expresses each variable \mathbf{y}_i ($i = 1, 2, \dots, m$) as a linear combination of latent factors \mathbf{f}_j ($j = 1, 2, \dots, p$), where p is the selected number of factors. So, for example for variable \mathbf{y}_1 the EFA model can be presented as

$$\mathbf{y}_1 - \bar{y}_1 = \lambda_{11}\mathbf{f}_1 + \lambda_{12}\mathbf{f}_2 + \dots + \lambda_{1p}\mathbf{f}_p + \varepsilon_1, \quad (4)$$

where the ~~residual error~~ term ε_1 accounts for the unique variance, m is the number of factors, \bar{y}_1 is the mean of the variable \mathbf{y}_1 , and λ_{ij} are the elements from the loading matrix $\boldsymbol{\lambda}$ and serve as weights to show how each factor \mathbf{f}_j is contributing for each variable \mathbf{y}_i (Morrison, 2005; Rencher and Christensen, 2012). [As for the PCA explained in the previous section, the Eq. \(4\) here describes the form of an EFA model based on literature, not the direct calculation in the algorithm. Thus, no scaling \(what e.g. the subtraction of the mean \$\bar{y}_1\$ from \$\mathbf{y}_1\$ in Eq. \(4\) essentially is\) is applied here.](#)

Different methods exist to calculate the factorization in EFA. In this study, principal axis factoring (hereafter pa-EFA) and maximum likelihood factor analysis (hereafter ml-EFA) were selected due to their suitability for our data (explained in more detail in section 3.3). In pa-EFA, the minimizable function F_1 can be presented as

$$F_1 = \sum_i \sum_j (S_{ij} - R_{ij})^2, \quad (5)$$

where S_{ij} is an element of the observed correlation matrix \mathbf{S} and R_{ij} is the element of the implied correlation matrix \mathbf{R} . Maximum likelihood factor analysis, on the other hand, minimizes the function F_2

$$F_2 = \sum_i \sum_j \frac{(S_{ij} - R_{ij})^2}{u_i^2 u_j^2}, \quad (6)$$

where the variances u_i and u_j for the variables i and j are considered. In other words, ml-EFA assigns less weight to the weaker correlations between the variables (de Winter and Dodou, 2012; Rencher and Christensen, 2012; Tabachnick and Fidell, 2014). In contrast to PCA, rotations are a recommended practice before interpreting the results in EFA, and the unrotated factor matrices are rarely useful (Osborne, 2014). Oblique Oblimin rotation was used to rotate the EFA factors. Orthogonal Varimax rotation was also tested, but as the orthogonality assumption for the factors is rather stringent for this type of chemical data, and it produced uninterpretable factors, those results are omitted. EFA was run in R Statistical Software with an addition of the psych-package (Revelle, 2018; R Core Team, 2019) and the factor scores were calculated as described above for PCA. The factors from EFA in the results are labelled as FE.

3.1.3 Positive Matrix Factorization (PMF)

PMF is a bilinear model and can be presented as

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} \quad (7)$$

where the original data matrix \mathbf{X} is approximated with matrices \mathbf{G} and \mathbf{F} , and \mathbf{E} is the residual matrix, i.e., the difference between observations in \mathbf{X} and the approximation \mathbf{GF} . After the factorization rank is defined by the user, Eq. (7) is solved iteratively in least squares sense. The values of \mathbf{G} and \mathbf{F} are constrained to be positive, and the object function Q is minimized (Paatero, 1997)

$$5 \quad Q = \sum_{i=1}^m \sum_{j=1}^n \left(\frac{E_{ij}}{\mu_{ij}} \right)^2, \quad (8)$$

The term μ_{ij} in Eq. (8) includes the measurement uncertainties for the observation matrix \mathbf{X} at time point i for ion j . Originally, μ was calculated as the standard deviations of \mathbf{X} , but other error types have also been used (Paatero and Tapper, 1994; Paatero, 1997; Yan et al., 2016). As apparent from Eq. (8), the measurement errors (μ_{ij}) act as weighting values for the data matrix. Thus, the chosen error scheme can have significant impact on the behaviour of Q .

10 To test this, different error schemes were investigated. The standard deviation values alone were not used as an error as the data includes fast concentration changes due to the sudden ignition of photo-oxidation which causes the standard deviations to be systematically too large. But as a reference, the standard error of the mean (the standard deviations of the ion traces divided by the square root of the number of observations, i.e., length of the ion time series) was used as an error for both [PTRPTR-MS](#) and AMS data. [It considers that measurements with less observations contain more uncertainty.](#) These error values are constant
 15 for each ion throughout the time series and do not change with signal intensity. This type of error is labelled here as static error. In addition, a minimum error estimate was applied, as suggested by Ulbrich et al., 2009). Determination of the minimum error for [PTRPTR-MS](#) is presented in the SI section S2.1 and for AMS in section S2.2.

Additionally, an error following the changes in ion concentration was constructed for [PTRPTR-MS](#) data by applying a local polynomial regression to smooth the ion time series (R-function loess, Cleveland et al., 1992). From the regression fit the
 20 residuals were calculated and the running standard deviation from the residuals was used as an error. Again, the minimum error was applied here. This error is referred hereafter as signal following error. For AMS, we also applied a [standard error that is frequently used by the AMS community. The standard AMS error consist of minimum error related duty cycle of the instrument and counting statistics following the Poisson distribution](#) (Allan et al., 2003; Ulbrich et al., 2009). [Poisson type error described in detail by Allan et al., 2003\), as this is frequently used by the AMS community.](#) Shortly, the [standard AMS error for signal \$I\$ Poisson error \$\mu_{ij}\$ for a single ion \$j\$](#) can be formulated as

$$25 \quad \mu_{ij}^{err} = \alpha \sqrt{\frac{I_o + I_c t_s}{t_s}}, \quad (9)$$

where α is an empirically determined constant (here $\alpha = 1.2$, generated by the AMS analysis software [PIKA](#), <http://cires1.colorado.edu/jimenez-group/ToFAMSResources/ToFSoftware/index.html>, last visit on 9.9.2019), [\$I_o\$ and \$I_c\$ are](#) the raw signal of the ion [of particle beam \$j\$ at time \$i\$](#) (ions s^{-1}) [for the chopper at open and closed -position, respectively.](#)
 30 and t_s is the sampling time [at a particular m/z channel](#) (s).

Examples of the used error values for [PTRPTR-MS](#) and AMS data are presented in SI at the end of sections S2.1 and S2.2, and the signal-to-noise ratios for different error matrices are reported in Sect. S2.3. [In contrast to the suggested best practice](#)

(Paatero and Hopke, 2003), we did not downweighed any ions in our data sets. This approach was used in order to give each SDRT equal starting point for the analysis, as e.g. for NMF or PCA similar downweighing is not possible, as we do not have any error estimates to calculate the signal-to-noise ratios in similar manner. However, to avoid misguiding the reader to omit recommended data pre-processing practice for PMF, we also tested PMF with downweighing. This, as expected, did not change our results significantly, but we acknowledge it should be indeed applied if aiming for a more detailed chemical interpretation of the PMF factors.

Often, constraining the values to be positive is not enough to produce a unique PMF solution for Eq. (7). This can be assessed by applying rotations, as in EFA and PCA. The rotations in PMF are controlled through the F_{peak} -parameter in which the changes produce new \mathbf{G} and \mathbf{F} matrices by holding the Q value approximately constant (Paatero et al., 2002). In this study, rotations with $F_{peak} = (-1, -0.5, 0, 0.5, 1)$ were tested. PMF analyses were conducted in Igor Pro 7 (WaveMetrics, Inc., Portland, Oregon) with the PMF Evaluation Tool (Ulbrich et al., 2009). The acquired results were further processed in R Statistical Software (R Core Team, 2019). The factors from PMF are labelled as FP in the results,

3.1.4 Non-negative Matrix Factorization (NMF)

Non-negative matrix factorization was introduced to the wider public after Lee and Seung presented their application of NMF to facial image database in Nature (Lee and Seung, 1999). The method has since gained popularity, and it has been used in various scientific fields including, e.g., gene array analysis (Kim and Tidor, 2003; Brunet et al., 2004). As in PMF, the NMF solution is constricted to positive values only to simplify the interpretation of the results, and in principle, both of these methods attempt to solve the same bilinear equation. In contrast to PMF, the algorithms in NMF do not require an error matrix as an input and it makes therefore no assumptions of the measurement error, and therefore we present NMF here as a separate method from PMF.

In general, the mathematical formulation of NMF is similar to the one presented for PMF in Eq. (7) and can be presented as $\mathbf{X} \sim \mathbf{WH}$,

where \mathbf{X} is the positive data matrix ($n \times m$) and \mathbf{W} and \mathbf{H} are the nonnegative matrices from the factorization with sizes $n \times k_p$ and $k_p \times m$, respectively, and p is the desired factorization rank. (Brunet et al., 2004). The value of k is equivalent to the selected factorization rank p . Multiple algorithms to calculate NMF exists (Lee and Seung, 2001). Here, we present results from the method described by Lee and Seung, 2001; Brunet et al., 2004), as this created the best fit to the data. The matrices \mathbf{W} and \mathbf{H} are randomly initialized, and are updated with the formula given by (Brunet et al., 2004)

$$H_{au} \leftarrow H_{au} \frac{\sum_i W_{ia} X_{iu} / (WH)_{iu}}{\sum_{k_p} W_{k_p a}} \quad (11)$$

and

$$W_{iu} \leftarrow H W_{iu} \frac{\sum_u H_{au} X_{iu} / (WH)_{iu}}{\sum_v H_{av}} \quad (12)$$

The NMF analysis was run in R Statistical Software with the NMF-package (Gaujoux and Seoighe, 2010; R Core Team, 2019). The factors from NMF are labelled as FN in the results.

3.1.5 Calculation of the contribution of an ion to a factor/component/cluster

From all these methods two factorization matrices (time series and factor contribution) can be produced at the end. In PMF and NMF, both factorization matrices are calculated simultaneously, whereas in EFA, PCA and PAM the factor/component time series are calculated after the main algorithm. The factor/component time series show the behaviour of each factor/component during the experiment while the contribution of the different variables to each factor/component (factor/component scores or factor profiles) can be interpreted as the chemical composition of each factor/component. To help the reader visualize the similarities and differences in the results between EFA, PCA, PMF, and NMF in this paper, we calculated the “total factor contribution” of each factor/component to each ion, i.e., how much each factor/component contributes to the signal of a single ion. For PMF and NMF, the values in the factorization matrices (**F** and **H**, respectively) were extracted for each ion and scaled with the sum over all factors for each ion. For EFA and PCA, the absolute values of the loadings were calculated for each ion in each factor/component and then scaled by the sum of all factor loadings. This approach allowed us to compare the division of the ions in each factor/component between the different methods. However, ~~this~~ type of approach conceals the information of the negative factor loading-values in EFA and PCA (which are included in the calculation of factor/component time series as weights), but instead visualizes the general contribution of an ion to a factor.

Negative factor loadings may have different interpretations. It may indicate that the compound has decreasing effect on the factor, i.e. it acts as a sink for the compounds with positive loading in the same factor. In chamber experiments, negative loading may also refer to decreasing concentration of the compound participating in chemical reactions if it acts as a precursor for other compounds in the same factor. One example of this is benzene, detected (as C₆H₇⁺) by PTR-MS. When inspecting the original loading values from EFA, for example, it has negative loading in FE1 (identified as later/slowly forming products), and positive loading in FE4 (identified as precursors from car exhaust/background). As benzene originates from the car exhaust, it contributes positively to FE4. However, as it oxidizes over the course of the experiment (thus has decreasing concentration), it has a strong correlation with oxidation products but appears negative in FE1, which mostly includes those later generation products.

3.2 Clustering methods

3.2.1 Partitioning Around Medoids (PAM)

Partitioning around medoids (PAM) or k-medoids is a clustering algorithm in which the dataset is broken into groups in which the objects or observations share similar properties in a way that object in a cluster are more similar to each other than to the objects in other clusters. The PAM algorithm is fully described elsewhere (Kaufman and Rousseeuw, 1990). Briefly, PAM minimizes the distances between the points and the centre of the cluster (i.e., the medoid) which, in turn, describes the characteristics of the cluster. The distance matrix (often also referred to as dissimilarity matrix) from the observed data can be calculated in many ways. Here, the data was first standardized by subtracting the mean of each ion over the time series and

scaling each ion with the standard deviations of the ions. Then, the Euclidean distances (Rencher and Christensen, 2012) were calculated between the ions before providing the distance matrix for PAM. The selection of suitable distance metrics can be challenging and depends on the application and the data. For example, Äijälä et al. (2017) tested four different metrics in their study of pollution events. In our study, also two other distance metrics were tested; Manhattan distance (e.g., Pandit and Gupta, 2011) and correlation-based distance metric. The results, however, were similar to those acquired with Euclidean distances, and therefore not shown here. The clustering was performed in R Statistical Software applying the factoextra- and cluster-packages (Kassambara and Mundt, 2017; Maechler et al., 2018; R Core Team, 2019). Clusters are labelled as CL in the results.

Often clustering is applied to cluster the observations in the data (e.g. samples, time points). Here, we have applied the clustering to the variables instead, to group similarly behaving chemical compounds together. This means our calculated distance matrix provides the distances between the variables (i.e., ions), and the centre of the cluster is the “characteristic” ion for that specific cluster. The larger the distances, the “farther apart” the ions are, and ions with shorter distances should be assigned to the same cluster. There are several clustering methods especially meant for clustering of variables (Vigneau, 2016). The time series for clusters are calculated by summing the concentrations of the compounds in the specific cluster. The interpretation of the results from cluster analysis slightly differs from the interpretation of the results of the other SDRTs. Due to the nature of cluster analysis in general (except fuzzy clustering, see e.g., Kaufman and Rousseeuw, 1990), the variables (here ions) are strictly divided between the clusters whereas for the other SDRTs presented in this study, one ion may have different weighing parameters for different factors/components. Depending on the aim of the study and the type of the data, this property of cluster analysis may be considered either as an advantage or disadvantage. [One obvious advantage of cluster analysis \(or hard division techniques in general\) is computational time, especially if analysing long ambient data sets. For laboratory measurements, this most likely is not an issue. Hard division techniques have also been shown to work efficiently for VOC measurements when distinguishing between different coffee types \(espresso capsules\), where strict separation between clusters is needed, as shown in Sánchez-López et al., 2014. For source apportionments studies, where one variable might originate from multiple sources, cluster analysis using hard division technique is probably not as suitable, as softer division techniques, which can assign one variable to multiple sources/factors.](#)

3.3 Determining the number of factors/components/clusters

One of the most difficult tasks in dimension reduction is the choice for the new dimensions of the data. For EFA and PCA, there exist multiple different methods determining the suitable factor and components number. However, these are often more guidelines than strict rules when handling measurement data as the processes creating the compounds which were measured can be somewhat unpredictable at times. Additionally, as EFA and PCA were originally developed for normally distributed data, tests for determining the number of factors may be influenced if the criterion of normality is not met. Furthermore, the existing tests to investigate multivariate normality are often oversensitive, e.g., for outliers (Korkmaz et al., 2014), which may influence the results. The analysis results from EFA and PCA, however, can be reasonably interpreted despite the data

distribution, as the normality of the data mainly enhances the outcome and is not stated as a strict requirement (Tabachnick and Fidell, 2014). In addition, the two calculation methods selected for EFA were used as they are supposed to be more suitable for non-normally distributed data. ml-EFA is rather insensitive to changes in data distribution (Fuller and Hemmerle, 1966), whereas pa-EFA is actually suggested to be more efficient if the normality condition is not met (Fabrigar et al., 1999). In this study, the multivariate normality of the data was nonetheless investigated, and results are reported in the SI material.

For EFA and EVD-PCA, we used the scree test first introduced by (Cattell, 1966), Kaiser criterion (Kaiser, 1960) and parallel analysis (Horn, 1965) to investigate the suitable number of factors/components. In the scree test, the factor number is estimated by plotting the acquired eigenvalues (or explained variance) as a function of factor number (see e.g. Fig. 1c). A steep decrease or inflection point indicates the maximum number of usable factors. The Kaiser criterion suggest discharging all factors that have eigenvalues less than 1 (see, e.g., Fig. 1d). In parallel analysis, an artificial data set is created, and the eigenvalues are compared to the eigenvalues of the real data. Here, we created the artificial data set for parallel analysis by resampling the actual measurement data by randomizing across rows as suggested by Ruscio and Roche, 2012). For SVD-PCA, the inflection point can be inspected, e.g., from a plot where the explained variance is plotted as a function of component number (see, e.g., Fig S11 in SI section S3.1). In addition, for EFA, we calculated the standardized root mean residuals (SRMR; Hu and Bentler, 1998) and empirical Bayesian information criteria (BIC; Schwarz, 1978) -values. [These metrics measure slightly different properties of the model. BIC is a comparative measure of the fit, balancing between increased likelihood of the model and a penalty term for number of parameters. The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. See S3.2 for more details.](#) A steep decrease in the SRMR values could indicate the number of factors similarly to the scree-test with eigenvalues. From BIC, the minimum value suggests the best fitting model. It should be noted, however, that these methods may suggest slightly different number of factors/components. In addition, many statistical tests are often oversensitive if the data is not completely normally distributed (Ghasemi and Zahediasl, 2012), even if large sample sizes might improve test performance, and therefore, the final decision of the number of factors should be made after evaluating the interpretability of the results.

The suitable number of clusters for PAM was investigated with the Total Within Sum of Squares (TWSS; e.g. Syakur et al., 2018) and Gap statistics (see, e.g., Fig. 1e and 1f). Within-cluster sum of squares is a variability measure for the observations within a cluster, and for compact clusters the values are smaller as the variability within the cluster is smaller. By calculating the TWSS, preliminary guidelines for the number of clusters can be derived by inspecting the inflection point of the TWSS versus number of clusters graph (often referred as the “Elbow method”). In Gap statistics, described in detail, e.g., by Tibshirani et al., 2001), the theoretically most suitable number of clusters is determined from either the maximum value of the statistics or in a way that the smallest number of clusters is selected where the gap statistics is within one standard deviation of the gap statistics of the next cluster number.

Such straightforward statistical tests are not available for PMF, but one possible option is to inspect the relation between Q and $Q_{expected}$. Ideally, the value of $Q_{expected}$ corresponds to the degrees of freedom in the data (Paatero and Tapper, 1993; Paatero et al., 2002) and when $Q/Q_{expected}$ (hereafter Q/Q_{exp}) is plotted against the factorization rank, an inflection point may be notable

and addition of factors does not significantly change the minimum value of Q/Q_{exp} (Seinfeld and Pandis, 2016). It should be noted, however, that even if the Q/Q_{exp} summed over all ions and time steps is low, the corresponding values of individual ions may still either be rather large or very small, thus compensating each other, and resulting in a unreliably good overall Q/Q_{exp} value (Interactive comment Paatero to Yan, 2016). In addition, the used error scheme in PMF has a large impact on the Q -values. If the true measurement error was used, Q/Q_{exp} approaches a value of 1. If the chosen error values were larger than this, the Q/Q_{exp} values will approach a final value smaller than 1. Note that the shape of the Q/Q_{exp} versus number of factors curve is not affected much by the chosen error scheme (see, e.g., Fig. 2b). Therefore, this method should be rather used as a first suggestion than a strict criterion. A more empirical method for determining the number of factors and interpretation of them exists when investigating ambient AMS data. The acquired factor mass spectra from PMF can be compared to spectra from known sources (Zhang et al., 2011). The time series of these identified factors are then compared to tracer compounds for these factors measured with other instruments (e.g. NO_x for traffic emissions, black carbon for burning events). If several factors correlate with the same tracers it is very likely that too many factors have been chosen. An extensive data base of factor spectra exists for AMS data and it is maintained by the community (<http://cires1.colorado.edu/jimenez-group/TDPBMSsd/>, last visit on 9.9.2019). The PMF evaluation tool for Igor Pro used in this study also provides other indexes, including the “explained variance/fraction of the signal”, which is shortly discussed in section 3.4.

Several approaches exist for NMF for selecting the factorization rank p , but the choice of which method to use is not straightforward (Yan et al., 2019). Brunet et al., 2004) suggested to select the factorization rank based on the decrease in the cophenetic correlation coefficient (CCC), i.e., at the first value of p where the coefficient decreases (see, e.g., Fig. 2a). In addition, we investigated the cost function that approximates the quality factorization as a function of the factorization rank p . For the brunet-algorithm that we applied in this study, this cost function is the divergence between data matrix X and approximation WH (see Eq. (3) in Lee and Seung (2001)). Other proposed indicator for optimal number of factors is the variation in the residual sum of squares (RSS) in which an inflection point may be detectable and suggest the factorization rank (Hutchins et al., 2008).

3.4 Determining the “goodness of fit”

When analysing the datasets, we realized found out saw that all of the factorization methods in this study are sensitive to even only small changes in the data. In order to cross-validate ~~whether~~ the calculated factorization and approximate the uncertainty in the factors is real and was not achieved only by change, 20 resamples of the measurement data were created with bootstrap-type sampling (Efron and Tibshirani, 1986), i.e., sampling with replacement from the original data. The resamples were formed by taking random samples (by row) from the measurement data with replicates allowed while preserving the structure of the time series. The different methods were then applied to the resamples to validate if the factorization created from the original measurement data was real and the created factorization was robust enough to maintain the achieved factor structure even if minor changes would appear in the data. Simplified, this variation in the factorization for the bootstrap type resamples can be understood as an uncertainty for the factorization results. If we had true replicates of the

data set, a similar approach could be used, as in theory the same, repeated experiments with similar chemistry should include the same factors, and the occurring variation in the factorization illustrate the uncertainties in the factorization.

In addition to the cross-validation of the factorization, the results should be evaluated in a way that we are able to justify how well the factors/components/clusters represent the original data and the underlying information. Often in studies where either EFA or PCA has been used, Explained Variance (EV) is reported for the solution. In principle, the EV could also be used as a guide when selecting the number of factors by selecting factors until EV reaches “appropriate” value or does not change drastically when more factors are added. In PCA, EV for each component is calculated by dividing the eigenvalue of each component by the sum of the eigenvalues. The sum of EVs for all $n-1$ components (n is the number of variables in the data) equals 1. In EFA the EV for the factor k (with p factors in total) can be calculated by

$$10 \quad \mathbf{EV}_k = \frac{\sum_{i=1}^n (\lambda_{ik}^2)}{\sum_{i=1}^n (\sum_{j=1}^p (\lambda_{ij}^2) + \text{diag}(\mathbf{S}-\mathbf{R}))}, \quad (13)$$

where λ_{ij} is an element from the loading matrix, \mathbf{S} is the original correlation matrix, and \mathbf{R} is the reconstructed correlation matrix ($\mathbf{R} = \boldsymbol{\lambda}\boldsymbol{\lambda}^T$) (Revelle, 2018). Depending on the algorithm used to calculate EFA, the calculation of EV may vary. In PMF, the calculation of EV is not possible this way, as PMF factorizes the data matrix instead of the correlation matrix. Instead, for PMF there is a possibility to calculate, e.g., the “explained fraction of the signal” for the reconstructed factor model. This can be calculated by comparing the original total time series (sum of the data columns, i.e., individual ion time series) to the reconstructed one by

$$15 \quad \mathbf{Frac} = \frac{\text{mean}(\sum_{j=1}^n (x_{ij}^*))}{\text{mean}(\sum_{j=1}^n (x_{ij}))}, \quad (14)$$

where x_{ij}^* is the element from the recalculated data matrix $\mathbf{X}^* = \mathbf{G}\mathbf{F}$ (see Eq. (7)), x_{ij} is an element from the original data \mathbf{X} , and n is the number of columns (variables, ions) in the data. The disadvantage of this method is the use of the mean. If the signal is both over- and underestimated at different parts of the data, the explained fraction of signal is still very good even if the fit is not. For NMF, a similar index could be calculated. However, due to the differences between EFA/PCA, NMF/PMF, and PAM (uses a fundamentally different approach), the indexes calculated with Eq. (13) and (14) are not comparable between the methods, and therefore not presented here. Instead, we aim for more universal ways to compare the SDRTs.

For NMF and PMF, it is possible to back-calculate how well the created factorization can reproduce the information in the original data. This method is rather straightforward, as both factorization matrices from NMF and PMF are limited to positive values. This allows us to calculate the reconstructed total signal for NMF/PMF, which can be compared to the original total signal to produce residuals. For EFA and PCA, the calculation of the total signal is not possible from the created factorization in a similar fashion, as the acquired loading values (contribution of an ion to a factor/component) may be negative. Therefore, for EFA and PCA the reconstruction is possible only for the correlation matrix, as it is also the matrix that is factorized during the calculation process. This allows us to compare the original correlation matrix to the one produced by EFA or PCA in a similar manner as for the whole data in PMF and NMF. However, due to our large data size, the visualization of the residual correlation matrix is difficult, and instead we have calculated the mean and interquartile ranges (IQR: Q_3-Q_1) for the absolute

values of the residuals. The theoretical minimum value for the mean and IQR is 0, indicating perfect reconstruction, and the theoretical maximum value, i.e., poor reconstruction, is 1. For example, for a variable-pair having a correlation coefficient of 0.7, a mean absolute correlation residual of 0.02 and an IQR of 0.04, this would mean the model over- or underestimates the correlation by 2.86% ($= (0.02/0.7) * 100$). An IQR of 0.04 would mean that 50 % of all variable-pairs with correlation of 0.7 are within 5.7% ($= (0.04/0.7) * 100$) of the original value of 0.7.

A very important criterion for the quality of the factorization is the interpretability of the results. If the interpretation of the factors is impossible, the results are useless for the data analysis. Note that all methods presented in this paper are purely based on mathematics, and the “best” result is obtained by solving a computational problem not connected to the real processes in the chamber and instruments leading to the measured dataset. Thus, the user has to apply the available external information (e.g., about possible reaction products or if ions should be split between multiple factors) to validate the feasibility of a factorization result. But it is a fine line between applying this prior knowledge about the possible chemical and physical processes in the chamber to validate a factorization result and dismissing an unexpected feature discovered by the factorization method as “unphysical” and thus wrong. Applying more than one factorization method may be helpful to protect the user from dismissing unexpected results.

4 Results and Discussion

4.1 Gas phase composition from PTR-MS

4.1.1 EFA

Figure 1 shows results for the tests described in section 3.3. The eigenvalues and parallel analysis results for EFA are not shown, as the results were very similar to those acquired for PCA. Also, the factorization results from ml-EFA and pa-EFA were so similar, that only the results from ml-EFA are presented here. Fig. 1a shows the empirical BIC and Fig. 1b the SRMR values for factorization ranks ranging from 1 to 10. The minimum value in empirical BIC was achieved with 4 factors and the inflection point in SRMR also lies around 4 factors.

As all these tests suggest a 4-factor solution for PCA and EFA, we compared the factor time series and factor contribution for the 4-factor (Fig. 3) and 5-factor solution (Fig S9 in SI section S4.1) for EFA with Oblimin rotation. The additional FE5 seems to be a mixed factor with small concentration created from FE4 and FE2, instead of a new factor with different properties. The original loading values for the 4-factor solution are presented in Fig. S10 as a scatter plot.

The variation in the factors from the resamples is largest around the start of the photo-oxidation, as expected, when there are fast and large changes in the concentrations. The mean and IQR for the absolute values of residual correlations for the 4-factor solutions were 0.0109 and 0.0108, indicating good reconstruction.

In the following, we interpret the factors for all SDTR methods based on their characteristic factor time series shape and the identified compounds in the factors. An overview of this interpretation is given in Table 1. Based on the shapes of the factor

time series, FE1 can be identified as an oxidation reaction product factor. It starts increasing slowly when the photo-oxidation starts, so these are either products from slow reactions, or multiple reactions steps are needed before these compounds are formed. FE2 is also an oxidation reaction product factor, but these are first (or early) generation products which rise quickly after photo-oxidation starts and are slowly removed by consecutive reactions as the photo-oxidation continues and/or by partitioning to the particle phase or chamber walls. FE3 is a precursor factor which shoots up during the α -pinene addition (slightly after $t = -50$ min) and is stable until the start of the photo-oxidation. Together with the factor mass spectrum which is dominated by signals at m/z 137 and 81, this is a clear indication that FE3 represents α -pinene in the chamber. Note that although proton transfer is a relatively soft ionisation technique, a certain amount of fragmentation of the mother molecule α -pinene (m/z 137) is observed showing fragments at, e.g., m/z 81 (Kari et al., 2018). FE4 seems to include some car exhaust VOCs and residue from the background. It has very low concentrations compared to the other factors. It decreases slightly throughout the whole experiment and seems not affected by the onset of photo-oxidation.

4.1.2 PCA

Figure 1c shows the eigenvalues as a function of component number for EVD-PCA with the results from parallel analysis. In Fig. 1d the eigenvalues for the first two components are omitted to show the changes with more components better. The blue line shows the Kaiser criterion (eigenvalue = 1). SVD-PCA ([when applied to scaled data matrix](#)) was not able to separate α -pinene as its own component, but instead created two factors which were dominated by the unreacted α -pinene and its fragments (see Fig. S12 in SI section S4.1). In addition, the unrotated solution included a large number of negative loadings, which complicated the interpretation of the components. [No improvement was achieved when SVD-PCA was applied to the data matrix without any scaling \(see. Fig. S13\).](#) Oblimin rotation was applied to create factors that could be interpreted in a physically more meaningful way, but the algorithm did not converge. So this is a case where the result of the factorization method is very difficult to interpret or even contrary to the available information (e.g. the α -pinene precursor behaviour). As additionally the underlying algorithm is struggling with the dataset (i.e., not converging), we will not discuss these results in detail here but rather focus on the EVD-PCA.

The number of components indicated by parallel analysis is 4 (Fig. 1c), but the eigenvalues decrease below 1 only with 10 components (Fig. 1d), indicating 9 components should be selected. However, the eigenvalues for the components 5-9 are rather close to the Kaiser limit (between 1.47 and 1.04, respectively), and therefore the 4-component solution was selected. In addition, the “knee” in the eigenvalues is around 4 or 5 components, but as for EFA, the addition of a fifth component did not create a new component with different properties but mixed properties of the previous components.

Figure 4 shows the component time series and total contribution from EVD-PCA with Oblimin rotation [and the original loading values for the 4-component solution are presented in Fig. S14 as a scatter plot](#). Oblique rotation was used despite the orthogonality assumption of the components, as for true physical components the assumption of orthogonality is not that realistic either, as it would indicate that the chemical processes taking place in the chamber don't have any correlation between the different processes. Oblique rotations allow correlation between the components, meaning the detected ions in different

components interact with each other. For example, the decrease of α -pinene concentration is mostly caused by chemical processes which in return form other ions detected by PTR-MS. Additionally, there are multiple consecutive processes (reactions) at work simultaneously, so the correlation between the components is not a straightforward indicator of connected processes, but it is more realistic than no correlation at all.

5 The mean and IQR for the absolute values of the residuals of the correlations were 0.0116 and 0.0107, respectively. Compared to the EFA solution with 4 factors, the residuals are slightly larger. The total contribution of compounds to each factor is very similar for EFA and EVD-PCA (Figs. 3b and 4b, or Figs. S10 and S14), which agrees with the very similar factor/component time series in general in the Figs. 3a and 4a. The interpretation of the components CO1, CO2, CO3, and CO4 is therefore the same as above for the EFA factors FE1, FE2, FE3, and FE4, respectively.

10 4.1.3 PAM

The test parameters TWSS and gap statistics for PAM are shown in Fig. 1e and 1f. The TWSS vs number of cluster values do not show a clear inflection point, but it could be roughly assigned between 3 and 5 clusters. There is no maximum value reached with gap statistics which indicates the theoretical number of clusters is 9, as there the gap statistic is within one standard deviation of the gap value in the 10-cluster solution. However, the increase of the gap value clearly slows down after 15 3 clusters. After careful evaluation, the 4-cluster solution is determined as most interpretable, and the cluster time series and distribution of the ions are shown in Fig. 5. Four clusters were selected, as the selection of only 3 clusters (Fig. S153 in SI section S4.1) is not enough to explain the variation in the data, as the addition of one cluster reveals new features. On the other hand, the 5-cluster (Fig. S164 in SI section S3.1) solution seems to split off an additional low-concentration cluster from CL4 (Fig. 5a), instead of showing a new distinct cluster. The distinction between the “more correct” solution with 4- or 5-clusters is not, however, straightforward, as CL5 (Fig. S164) could be interpreted as a car exhaust precursor cluster, as is CL4 in Fig. 20 5. Clustering statistics are presented in Table S24 (SI Sect. S4.1).

When comparing the shape of the cluster time series to the EFA and PCA results in Fig. 3 and 4, the results agree well. The largest difference appears in the CL4, which has larger concentrations in PAM compared to FE4 acquired from EFA and CO4 from PCA. The shapes of FE4 and CL4 are also slightly different, as CL4 has a small decrease in the concentration at = 0 min 25 whereas FE4 is barely affected. However, comparing the actual concentrations between these methods (EFA/PCA and PAM) may be misleading, as in EFA and PCA, the acquired loading values are used as weights when calculating the factor time series whereas in PAM the time series of the clusters are calculated as a direct sum of the cluster compounds, as explained in chapter 3.2.1.

Dichotomized loadings (for each ion: 1 for factor with largest loading, 0 for the other factors) for EFA were tested to see if 30 then the results agree better with those from PAM as in PAM there are no loading values, meaning an ion either is in a cluster or not. With dichotomized EFA loadings we make the same assumption: one ion is classified to one factor only, and therefore stemming from only one source or source process. Figure S175 (SI Sect. S4.1) shows the results from dichotomized EFA. When compared to PAM (Fig. 5), the factor/cluster concentrations agree well, but there are clear differences in the ion

distribution. EFA classifies the weak ions with low concentration to the product factors (FE1, FE2), whereas PAM assigns them to the background/precursors cluster (CL4).

4.1.4 NMF

Figure 2a shows the [divergence of the cost function \$D\(X||WH\)\$](#) and CCC for factorization ranks from 2 to 10 for NMF.

5 The CCC has a first decrease in the values at the rank 4 and the [residual sum of squares \$D\(X||WH\)\$](#) shows an inflection point around the ranks 4-5. Figure 6 shows the factor time series and total contribution for the NMF with factorization rank 5. Five factors were selected, [even though CCC suggest only 4 factors](#), as the addition of one factor to the 4-factor solution (Fig. S186 in Sect. S4.2) did add a new feature to the solution in contrast to the SDRTs presented above. FN2 in the 4-factor solution decreases drastically between $t = -50$ and $t = 0$, indicating it might include background ions, but on the other hand, it also peaks right after $t = 0$, indicating oxidation products also contribute to that factor. These mixed properties in the factor FN2 indicate more factors are needed and indeed in the 5-factor solution this contradictory behaviour no longer occurs.

Similar to the results shown above, the range in the factor time series for the bootstrap replicates is larger when the factors exhibit fast changes in the concentration (Fig. 6a). In addition, FN3 from the real measurement data has a lower maximum concentration when compared to the bootstrap replicates. This indicates NMF is rather sensitive to the small changes in the data and only a few deviant observations present in data but not in majority of the resamples can cause this kind of discrepancy. Factors FN1-4 seem to correspond to the same factors found with EFA, PCA, and PAM (Fig. 2, 3 and 4), and especially α -pinene is clearly assigned to the same factor (FE3/CO3/CL3/FN3) in all the used methods. FN5 in NMF, however, has properties that were not detected (or separated from others) with EFA, PCA, or PAM even if more factors were added. This new factor could be interpreted also as oxidation product factor, but as it increases slower and decreases later than the early product factor (FN2), it mostly includes intermediate products. These are most likely compounds which are formed through (multiple) reactions and consumed in further oxidation reactions.

By recalculating the data matrix, \mathbf{X} , with the original factorization matrices \mathbf{W} and \mathbf{H} , we can inspect how well it has been reproduced. Here, the total signal (total time series) is then calculated by summing all ions for each time step for the original data matrix and the reconstructed data matrix. The differences between the original total signal and the one produced by NMF (i.e., the residuals) were smaller than 10^{-10} , indicating a good [mathematical](#) reconstruction. The boxplot of the residuals with 4 and 5 factors is shown in Fig. S197 (SI Sect. S4.2)

4.1.5 PMF

The acquired Q/Q_{exp} for different factorization ranks in PMF with the constant error scheme are presented in Fig. 2b. The values are the minimum values from all possible solutions with f_{peak} values from -1 to 1 by a step of 0.5. The Q/Q_{exp} were at the minimum at $f_{peak} = 0$ with the number of factors (1-10) tested. We notice that the values are slightly smaller in general when using the signal following error as the absolute values of the errors in this error scheme are significantly larger around fast changes than in the static error scheme and thus decreasing the observed Q values (see Fig. S4 for the different error

schemes). Values for the signal following error decrease slightly below 1 (0.88) for the 5-factor solution, whereas with the static error they stay above 1.91. After careful evaluation of the results with different number of factors, the solution with 5 factors (Fig. 7, $f_{\text{peak}} = 0$) was selected to be presented and interpreted here. The solutions with fewer factors were inconclusive, and the addition of a fifth factor did add a new feature. The results with 4 factors are shown in the SI material (Sect. S4.2, Fig. S2018). In the 5-factor solution, the solid lines in the time series are the results for the measured data, and the shaded areas show the ranges for the bootstrap resamples.

For the static error case, the factorization from resamples agree well with the ones from measurement data. For the signal following error (Fig. 7c-d) the differences are significantly larger, and for example FP5 has a larger peak concentration than in any of the resamples. This is most likely caused by few deviant values in the data which are not present in the resamples, and thus creating smaller peak concentration for FP5 for the resampled data. Resampled data includes more sudden changes due to added and/or missing data rows, thus causing PMF to perform poorer. In addition, the other variations in the resampled ion time series may cause ion contributions (especially those originally assigned in-between factors) to shift slightly from FP5 to FP2, as for FP2 the values in the time series are higher in the resamples compared to the original data. This difference between the error schemes is caused by the error values themselves. For the signal following error, the factorization is more “precise” (less wiggly factors), but even small shifts in the data (bootstrap resamples) distort the factorization more than in the static error case.

When comparing the results for the measurement data with different error schemes in Fig. 7, we note that the α -pinene precursor factor is slightly less pronounced with the signal following error, i.e., the solving algorithm assumes these fast changes are not “real” but rather outliers. This is caused by the used error scheme, where errors are larger for the fast changes in the data (Fig. S4b). ~~In For~~ ambient data not measured at instant proximity of strong emission sources, for which PMF is often used~~was developed~~, this type of error is beneficial as there the fast changes are mostly more likely to be noise or instrument malfunctions (excluding, for example, sudden primary emission plumes), and we are more interested of the long-term changes instead. For laboratory data, where large changes are often caused by rapid changes in actual experimental conditions, e.g. due to injecting α -pinene or turning the UV lights on, the static type of error is most likely preferable. Using Usage of the static error scheme helps to avoid ~~underweighting-overcorrecting-~~ intentional (large) changes in experimental conditions and confusing them with real variation taking place during the experiment and typically being much less pronounced.

Figure 8 shows how well the original data matrix can be reproduced with the created factorization matrices. The residuals for the static error are generally larger as most of them are in the range 0 ± 0.5 (for signal following error 0 ± 0.15), but there are much larger “outlier” values for the signal following error. This is due to the structure of the signal following error, which is larger during the fast changes in the data, as shown in Fig. S45b (SI Sect. S1.4). For the static error- the residuals vary more throughout the whole data, whereas for the signal following error the residuals are smaller, but a few rather large values appear at the start of the photo-oxidation, as seen in Fig. 8b. This highlights the role of the selected error values in PMF, which act as weights for the data. Smaller error value means that the corresponding Q value at this time will be much larger and an

improvement of the model at this part of the data will have a big impact on the optimisation value. This means that the error values can be used to emphasise certain parts of the dataset which otherwise would not be recovered very well by PMF. Note that this is a key difference to NMF where no error-based weighting of the data is done.

4.1.6 Comparing the SDRTs applied to gas phase composition

5 Table 1 summarizes the acquired results from different SDRTs for the gas phase composition data measured with [PTRPTR-MS](#), and [Figs. S21 and S22 in SI section S5 show separate factor contributions for each of the SDRTs](#). Comparison of the total factor contribution for some selected compounds for the 4 factors from EFA, PCA and PAM are shown in Fig. 9. We note that the differences are very minor between EFA and PCA, and hardly visible in the coloured bars. When compared to PAM, we see that e.g. acetaldehyde and methyl ketene are assigned to the red cluster (CL2), which is also dominating in EFA and PCA
10 (FE2, CO2). Figure 10 shows the same compounds for NMF and PMF. There, the largest difference is between the 2 oxidation product factors, coloured as black (late oxidation products, factor1 in Table 1) and red (fast oxidation products, factor2 in Table 1). For the selected compounds, NMF has more weight assigned to the fast forming products than PMF. In addition, PMF assigns much more weight to the intermediate oxidation product factor (pink) for some of the compounds.

The factorization acquired from PMF agrees well with the factorization from NMF when comparing the factor time series, as
15 expected since the methods are rather similar. Comparison of the concentrations of the factors between PMF and NMF directly is not exact, as these methods have different weighting between the produced factorization matrices due to the different solving algorithms. The largest difference is the early product factor, FN/FP 2. In NMF (Fig. 6), this factor (FN2) increases from 0 ppb to 30 ppb very fast at $t = 0$ min, then it decreases rapidly to just above 20 ppb and continues to decrease almost linearly towards 10 ppb. In PMF, this factor (FP2, Fig. 7) has a similar increase at $t = 0$ min, but it decreases exponentially instead of
20 the fast drop and constant decrease present in NMF solution. The different slope has direct implications for the interpretation of the factor. A faster decrease is interpreted as a faster removal/destruction process for ions classified into this factor. This is typically related to reaction speeds or to how far along a product is in the chain of oxidation reactions. When comparing the total contribution of FN2 in NMF and FP2 in PMF, in NMF ions with m/z 90-100 have much more contribution to FN2, whereas in PMF these ions seem to be assigned to FP1 instead. Otherwise the factors agree well with those acquired from
25 NMF, and their interpretation is therefore similar: 3 oxidation product factors (FP1, FP2, and FP5), 1 background/car exhaust precursor factor (FP4), and 1 α -pinene injection factor (FP3).

Another important difference between NMF and PMF is the relation between the factors at the end of the experiment. In PMF, at the end everything is shifted to FP1 (later generation oxidation products) and the other factors decrease to 0, whereas in NMF there still is contribution from the other oxidation product factors FN2 and FN5 in addition to FN1. More fundamental
30 studying of the algorithms for both PMF and NMF is needed to explain this behaviour.

The factorization acquired with EFA, PCA and PAM is more robust compared to NMF and PMF when inspecting the bootstrap ranges in the top panels in Figs. 3, 4, 5, 6 and 7. This may be explained with the different number of factors (4 or 5) as with more factors, one factor includes less (strongly) contributing ions, which causes factorization to vary more when the data is

different. But most of the differences between these SDRTs are still explained by the methods themselves and the solving algorithms. PMF and NMF are more sensitive to small changes in the data, whereas EFA, PCA and PAM succeed more reproducibly in finding larger structures and changes in the data.

5 Addition of a fifth factor to EFA, PCA and PAM did not add a factor showing a new feature, as it did in NMF and PMF, but a sub-factor. This sub-factor has very low concentration, but if inspected separately (not shown), it peaks around $t = 0$ min similar to the factor2. This means, instead of adding a factor consisting of intermediate oxidation products (as in NMF and PMF), the added factor is another early product (or background) factor. This is also caused by the difference in the methods, as these 3 SDTRs (EFA, PCA, and PAM) concentrate more on the fast changes (which take place here at $t = 0$ min), whereas NMF and PMF focus more on slow changes. This is one example where the chosen method (EFA/PCA or NMF/PMF) has a
10 direct impact on the interpretation of the data. For understanding the chemical processes in the experiment, the existence of 2 or 3 oxidation product factors is of great importance.

Factor4 has different behaviour in the time series in EFA and PCA compared to NMF, PMF and PAM. In the latter SDRTs this factor starts decreasing immediately and the concentration drops through the whole experiment, implying it is affected by car exhaust precursors that are oxidized, and the products are assigned to other factors later on. In EFA and PCA, this factor
15 has small and rather stable concentration over the time series (also in addition to a small contribution to the total signal), suggesting it could consist of background compounds present throughout the whole experiment. Without exact identification of all the compounds present in these factors (which is out of the scope of this study) , it is hard to say if this difference is real or if it is related to the different calculation of the SDRTs. We provide more details of the comparison between the factors from different SDRTs in Sect. S5 in SI material.

20 4.2 Particle phase composition from AMS

4.2.1 EFA and PCA

As the AMS data from the experiment includes only one observation about every 10 minutes, the data has ~~many~~ more variables (compounds) than observations. This causes problems for EFA and EVD-PCA, as those methods are based on the correlation matrix, which will not be positive definite due to the small number of observations (rows) compared to the number
25 of variables (columns). In EVD-PCA, the second step is to calculate the eigenvalues which in this case may also be negative and result in a non-interpretable outcome. With this type of data, the results of EFA are also sensitive to the used algorithm. The calculation in ml-EFA did not converge at all, but pa-EFA was able to produce results. Due to these restrictions in the calculation process, the results from EFA and PCA are only briefly discussed below, and example figures can be found in the SI material (Sect. S6.1). In addition, due to the very small data size, the bootstrap-type resampling of the data has a too drastic
30 effect on the data structure to validate the repeatability of the factorization and is therefore not applied for any of the SDRTs. Figure S23+9a and b (SI material Sect. S6.1) shows the results for the tests investigating the correct number of factors and components pa-EFA and EVD-PCA. For EFA in Fig. S23+9a, the empirical BIC reaches a minimum value with 4 factors and

the inflection point in SRMR is at 4 factors. For EVD-PCA, however, parallel analysis (Fig. S23+9c) suggests only 1 component, mainly indicating the data is not suitable for PCA at all. The eigenvalues also do not reach 1 (Kaiser criterion) with up to 10 components tested. The eigenvalues for EFA (not shown) reached 1 for the 6-factor solution, and parallel analysis results (not shown) indicated to select only 1 factor. The differences in these test results is mostly caused by the computational issues mentioned above. Indeed, neither EFA nor PCA (SVD or EVD) were able to separate more than two factors/components from the data, when 2-5 factors/components were tested (see e.g. Figs. S240 and S25+ in Sect. 6.1). While a 2-factor solution could be correct in principle, it seems unlikely for the investigated system. The particle phase is constantly formed by low volatile gaseous compounds condensing. As shown above, the gas phase composition changes constantly as compounds are produced and consumed. Thus, it is highly unlikely that during the 4 h of chemical reactions in the chamber the same mix of low volatile compounds is present and condenses onto the particles.

4.2.2 PAM

No clear inflection point is visible in the TWSS-plot in Fig. S23+9e; the value decreases when increasing the cluster number with a small “bump” at 4 clusters. The gab statistics (Fig. S23+9f) does not reach a maximum value, and it also does not reach the other criteria explained in Sect. 3.3. The inconclusiveness of the tests results may be caused by different reasons, and to investigate this further, PAM was conducted with 2 – 5 clusters and the results are shown in the SI material (Figs. S262 and S273). Increasing the number of clusters from 3 (Fig. S26b2) upwards adds clusters with extremely small concentrations and a similar time series shape to the previously found clusters. The very similar shape of the time series of the clusters suggests that only one type of SOA particles was formed quickly after the start of photo-oxidation, and that the chemical composition changed only marginally. Again, this seems unlikely [for the investigated system](#).

The inability of PAM to identify multiple SOA particle types most likely lies in the method itself. Each variable (ion) is assigned to one cluster and cannot be spread over multiple clusters. However, it is well known that AMS applies a “hard” ionisation technique. Thus, a high degree of fragmentation is expected and indeed, most carboxylic acids, for example, are detected as CO_2^+ (m/z 44). This means, a highly oxidised organic acid, formed late in the experiment after multiple steps of oxidation will be detected at the same variable (ion) as a different acid formed much earlier. Due to the “one variable-one cluster” method, PAM is incapable of resolving this information in the data. While EFA and PCA could still be used if the data matrix is suitable (i.e., more rows than columns), PAM is unsuitable for this AMS dataset or generally datasets where variables have strong contributions from more than one source.

4.2.3 NMF

The $\frac{D(X||WH)}{RSS}$ has an inflection point at factorization rank 4 and CCC shows the first decrease in the values with 4 factors, as shown in Fig. 11a. We selected the 4-factor solution for the detailed interpretation. We discuss additional reasons why the 4-factor solution should be selected in SI Sect. S6.2. The factor time series are shown in Fig. 12a. Fig. 12b shows the original ion-to-factor contributions from NMF without any scaling. The total factor contribution plots are omitted, as we don't have

PCA/EFA results to compare. The delay in the time series after $t = 0$ (before the factors starts increasing/decreasing) is most likely caused by the small time resolution (10 min) of the data. The residuals were on the same order of magnitude as for the PTR-MS data, indicating again very good reconstruction of the original signal.

The mass spectrum of FN1 is dominated by the $C_nH_{2n+1}^+$ and $C_nH_{2n-1}^+$ ion series, conforming to the typical features of combustion-related primary organic aerosol and thus it is interpreted as a hydrocarbon organic aerosol (HOA) -factor. FN1 was originated from car exhaust, as it already appears before $t = 0$ min. FN1 increases slightly at the start of the photooxidation. The increase is partly attributed to the new formation of the HOA component when the HOA type compounds in the hot exhaust gas was introduced to the chamber and contains marker ions associated with HOA (e.g. m/z 57, Zhang et al., 2005) and 59) condensed again in a cooler chamber. Meanwhile, we can't rule out the possibility that HOA has been produced as a minor product after the photooxidation reaction was enabled in this study. FN2 can be interpreted as α -pinene secondary organic aerosol (SOA) derived semi-volatile oxygenated organic aerosol (α P-SOA-SVOOA) after we carefully compared the factor mass spectra with pure α -pinene experiments conducted at similar settings reported by Kari et al., 2019b). In addition, FN2 is characterized by the prominent peak as m/z 43. The mass spectra of FN2 and FN4 are rather similar, but FN4 has more contribution from m/z 44, a marker of oxygenated organic aerosol (Zhang et al., 2005) and thus it is identified as an α P-SOA-LVOOA factor. The FN3 was appointed as a mixed LVOOA. Except for the high peak at m/z 44 in the FN3 mass spectrum, its time series is also consistent with the SOA formation in the mixed α -pinene/car exhaust SOA experiments conducted in similar settings (Kari et al., 2019b) and thus it is identified as a mixed LVOOA factor stemming from later generation oxidation products. Summary of the generated factors from NMF can be found in Table 2.

4.2.4 PMF

The O/O_{exp} for the two error schemes are shown in Fig. 11b. Neither of the error schemes show a clear inflection point. Examples of behaviour of the errors as a time series are shown in Fig. S7 (SI Sect. S2.2). With the standard AMS error, the O/O_{exp} values do not reach 1 (with 10 factors $O/O_{exp} = 1.76$) whereas with the static error the values decrease below 1 for 7 factors. The solutions with 2-5 factors were inspected, and the 2-factor solution (Fig. 13) is presented here as the most interpretable one (summarized in Table 2). The primary OA factor, separated by NMF (Fig. 12, FN2) was only found if using 4 factors and the static error scheme in PMF (see SI Sect. S6.2, Fig. S31a-b). However, interpretation of the time series for that solution was found to be very difficult due to the extreme anticorrelation between the time series, and thus the 2-factor solution was selected. The two factors were interpreted as SVOOA and LVOOA. In addition, the largest relative decrease in the O/O_{exp} was observed with the 2-factor solution.

The residuals for the standard AMS error were smaller, as shown in Fig. 14. This agrees with the analysis of the gas phase data set, where the residual for the signal following error (which has a similar profile in time as the standard AMS error) were generally smaller compared to the static error.

The signal following error, used for PTR-MS, was also tested for particle phase data. However, as this type of error showed very similar behaviour as a time series as the standard AMS error, and produced very similar outcome, those results are omitted from this manuscript.

The Q/Q_{exp} for the two error schemes are shown in Fig. 11b. Neither of the error schemes show a clear inflection point.

- 5 Examples of behaviour of the errors as a time series are shown in Fig. S7 (SI Sect. S2.2). With the Poisson type error, the Q/Q_{exp} values do not reach 1 (with 10 factors $Q/Q_{exp} = 1.76$) whereas with the static error the values decrease below 1 for 7 factors. The solutions with 3-5 factors were inspected, and the 4 factor solution with the static error (Fig. 13a-b) is presented here as the most interpretable one (summarized in Table 2). When comparing the distribution of ions in the factors with the static error scheme in Fig. 13b, the factors agree with those calculated with NMF in Fig. 12b
- 10 With only 3 factors the primary HOA factor was not separated with either of the error schemes (Fig. S26-S28 in SI Sect. S6.2), and the addition of the 5th factor did not add a factor with a new feature but instead created more sub factors (Fig. S27-S29). In the 4 factor case, with the Poisson style error, the primary HOA factor could not be separated (Fig. 13c-d), and instead a sub-factor, FP1, was created from the other three factors in the 4 factor solution. Nevertheless, the residuals for the Poisson style error were smaller, as shown in Fig. 14. This agrees with the analysis of the gas phase data set, where the residual for the signal following error (which has a similar profile in time as the Poisson style error) were generally smaller compared to the static error. The different factorization between the error schemes highlights not only the need for proper selection of the used error, but also interpretability of the factors. Computed reconstruction of the measured data is slightly better with the Poisson style error, but the interpretation of the factors is more difficult. Acquiring a balance between statistically good results and realistic factors might be challenging, and to achieve more robust results, testing different error schemes may be beneficial, especially
- 15 following error (which has a similar profile in time as the Poisson style error) were generally smaller compared to the static error. The different factorization between the error schemes highlights not only the need for proper selection of the used error, but also interpretability of the factors. Computed reconstruction of the measured data is slightly better with the Poisson style error, but the interpretation of the factors is more difficult. Acquiring a balance between statistically good results and realistic factors might be challenging, and to achieve more robust results, testing different error schemes may be beneficial, especially
- 20 if the data size is small.

4.2.5 Comparing the SDTRs applied to particle phase composition

Table 2 summarizes the acquired results from NMF and PMF for the particle phase composition data measured with AMS, and additional details are provided in the SI material (Sect. S5). PAM was not able to separate distinct clusters due to the inability of clustering techniques to classify an ion into multiple clusters. Comparing the relative factor spectra and fraction of signal from NMF and PMF with the static error (Figs. 12b and 13b), the distribution of ions is indeed similar between the LVOOA in the PMF solution and α P-SOA- the mixed LVOOA factor in the NMF solution, and also similar between the SVOOA in the PMF solution and the α P-SOA factor (integrated- α P-SOA-SVOOA and α P-SOA-LVOOA factors), respectively, but PMF has more contribution of high m/z compounds in the primary HOA factor (factor1). However, in the time series (Figs. 12a and 13a), more differences occur. For example, factor2 (α P-SOA-SVOOA) has a similar general structure (i.e., peak concentration around $t = 50$ min) in time series in both PMF and NMF, but in NMF the factor is smoother. The same applies to the other factors. When inspecting the individual ion time series in the original AMS data, most of them have rather “smooth” behaviour, similar to the factors from acquired from NMF. It seems, that PMF gives more weight to the background ions (with very small concentration) which do not have that clear structure in their time series, thus including more

25 signal from NMF and PMF with the static error (Figs. 12b and 13b), the distribution of ions is indeed similar between the LVOOA in the PMF solution and α P-SOA- the mixed LVOOA factor in the NMF solution, and also similar between the SVOOA in the PMF solution and the α P-SOA factor (integrated- α P-SOA-SVOOA and α P-SOA-LVOOA factors), respectively, but PMF has more contribution of high m/z compounds in the primary HOA factor (factor1). However, in the time series (Figs. 12a and 13a), more differences occur. For example, factor2 (α P-SOA-SVOOA) has a similar general structure (i.e., peak concentration around $t = 50$ min) in time series in both PMF and NMF, but in NMF the factor is smoother. The same applies to the other factors. When inspecting the individual ion time series in the original AMS data, most of them have rather “smooth” behaviour, similar to the factors from acquired from NMF. It seems, that PMF gives more weight to the background ions (with very small concentration) which do not have that clear structure in their time series, thus including more

30

of their behaviour in the final factors, [if the number of factors in PMF is increased from 2 \(see SI Sect. S6.2 Figs. S30-S32\)](#). Residuals from NMF reconstruction [\(with 4 factors\)](#) were over ten [orders of](#) magnitudes smaller (for NMF between $-1.3 \cdot 10^{-13}$ and $7.1 \cdot 10^{-14}$) than those from PMF (between -0.064 and 0.13 for [Poisson style error](#)~~the 2-factor solution with standard AMS error~~, see Fig. 14), indicating better reconstruction of the data with NMF. Most likely PMF struggles with the small data set, thus [not being able to recover all the factors found by NMF and construct reasonable time series for those factors \(see 4-factor solution in Fig. S31\)](#), ~~being unable to recover reasonable factors also as a time series~~ whereas NMF does not seem to be affected [by the data size](#). In addition, the weighing between the factorization matrices between NMF and PMF is different not only due to the error matrix that is given as a weight in PMF, but also because of the different solving algorithms for each method. This, on the other hand, assigns different emphasis between the matrices, possibly causing NMF to use more effort to reconstruct the data matrix with factor time series. [However, the reader should keep in mind that for detailed chemical analysis of such data sets, especially with PMF, downweighing is advisable. In addition, the replacement of very small negative numbers with very small positive numbers is not mandatory for PMF, as it can run with a few negative values into some extent. However, we did the replacement here, as the NMF algorithm used here, does require strictly positive input data. Acquiring a balance between statistically good results and realistic factors might be challenging, and to achieve more robust results, testing different error schemes may be beneficial, especially for](#) ~~if the data set with such small size. size is small.~~

4.3 Computational cost

To approximate the differences in the computational time between the different SDRTs, the methods were applied with 2-10 factors each, having 9 runs in total for each method. No rotations were applied (no rotation for EFA and PCA, $F_{\text{peak}} = 0$ for PMF), as the rotational methods between EFA/PCA and PMF are not directly comparable. Computation times include the calculation of the correlation matrix when needed and calculation of the factor time series for PAM, EFA and PCA (which is calculated outside the main algorithms), as described in Sect. 3.1.1. Three data sets with different sizes were tested, and the results are presented in Table 3. AMS includes the particle phase measurement data (size 26x306) presented in Sect. 4.2. and [PTR-MS](#) the gas phase composition data (size 300x133), which was analyzed in detail in Sect. 4.1. [PTR-MS*5](#) is a larger data set created from the gas phase composition data by duplicating the data rows 5 times (final size 1500x133). The computational times for NMF and PMF were clearly longer when comparing to the other SDRTs. This is not surprising, as PMF and NMF calculate both factorization matrices at the same time, whereas for the other SDRTs only the matrix presenting the contribution of ion to factor is found at first, and the time series of the factors/components/clusters are calculated afterwards. In addition, the PMF2 algorithm used through the PMF evaluation tool for Igor Pro does read and write text files for each PMF run, thus significantly increasing the computational time.

4.4 Summary of the SDRTs used in this study

The methods tested in this study have many similarities and many, fundamental or computational, differences between them. Though in literature, they are many times applied to similar problems. In this section we will summarize some of these properties.

- 5 EFA has a fundamental difference to the other methods as it is by definition a measurement model of a latent variable, i.e. the factor. (Osborne, 2014) whereas the other methods are basically describing the measured data with linear combinations of measured variables. The latent variables in EFA, i.e. the factors, cannot be directly measured, but instead, they are seen through the relationships they initiate in a set of Y variables, which are measured. In the other methods, in turn, the factors, components or clusters are calculated directly from the measured variables Y (Rencher and Christensen, 2012; Osborne, 2014).
- 10 The approach to data reduction in PCA is to create one or more summarizing variables from a larger set of measured variables by retaining as much as possible of the variation present in the original data set (e.g. Jolliffe, 2002). This is done by using a linear combination of a set of variables. The created summarizing variables are called components. The main idea of the PCA is to figure out how to optimize this process: the optimal number of components, the optimal choice of measured variables for each component, and the optimal weights when calculating the component scores.
- 15 The objective of cluster analysis is to divide the observations into homogeneous and distinct groups (Rencher and Christensen, 2012). Cluster analysis is a method where the aim is to discover unknown groups in the data, which are not known in advance. The goal of the clustering algorithm is to partition the observations into homogeneous groups by using some measure of similarity (or dissimilarity), such that the within-group similarities are large compared to the between-group similarities. The choice of the similarity measure can have a large effect on the result. One property of cluster analysis is that it will always
20 calculate clusters, even if there is no strong similarity present between the variables in the data (Wu, 2012). This should be noted when interpreting the results, especially if the user has no a priori information about the number of clusters.
NMF and PMF provide an alternative approach to the decomposition assuming that the data and the components are non-negative (Paatero and Tapper, 1994; Lee and Seung, 1999). Thus, all the features learned via NMF and PMF are additive; that is, they are adding together strictly positive features. PCA and EFA tend to group both positively correlated and negatively
25 correlated components together, as they are only looking for the correlations of variables (except SVD-PCA, which can be applied to the data matrix directly). On the other hand, NMF and PMF, by constricting W and H to positive values, find patterns with the same direction of correlation. Thus, NMF and PMF work well for modelling non-negative data with positive correlations. However, if the interest is not only in the positive effects, then PCA and EFA can provide more information of for the investigated system. Cluster analysis is suitable for classifying observations based on certain criteria. The researcher
30 can measure certain aspects of a group and divide them into specific categories using cluster analysis. However, this method is not suitable for data with variables which should show contributions from multiple factors/components (e.g. strongly fragmented signals in AMS data).

Factorization methods, including those used in this paper, operate on the fundamental assumption that the factor profiles (here factor mass spectra) are constant over the investigated period. Often, this has been interpreted in a way that chemical processes occurring in a chamber experiment or the atmosphere violate this assumption. However, this interpretation is based on a too narrow definition of what a factor represents. A factor can be seen as a direct (emission) source of compounds which changes its contribution to the whole signal (e.g. primary emissions from biomass burning as a fire develops and then dies). But a factor can also be interpreted as a group of compounds showing the same temporal behavior. If this group is released together as an emission or if the compounds are formed in the same ratio by some chemical process should not matter. In the latter case, it is important how wide the group is selected, i.e. if we group products of processes together for which the contribution changes with time. This means, choosing the optimal number of factors becomes even more important when chemical processes are occurring. EFA and PCA account for the chemistry happening in chamber measurements with negative loadings, as described above. Same factor can contain educts and products of a chemical process (e.g. oxidation), with the difference that their loadings are negative and positive, respectively.

Taking account of everything above, the most important thing to consider when selecting the SDRT is the interpretation. What are the features the researcher wants to deduct from the data, what are the properties of the data and how the data can answer the research questions. As we have shown in the results section, the methods provide quite similar reconstructions of the time series, but the interpretation of the steps leading to these are quite different. For example, comparing EFA factor FE4 in Fig. 3a and PMF factor FP4 in Fig. 7a reconstructions, they seem to show the same procedure but the first one includes both positive and negative effects and the second one consists only of positive effects.

5 Conclusions

The main objectives in this study were to a) investigate how different SDRTs perform for gas- and particle-phase composition data measured with mass spectrometers, b) how the interpretation of the factors changes depending on which SDRTs has been used and c) how well the SDTRs were able to resolve and classify the factors representing chemistry behind the investigated data set of photo-oxidation of car exhaust combined with α -pinene. We showed that EFA, PCA and PAM were able to identify 4 factors from the gas phase composition data, whereas NMF and PMF succeeded to separate one additional oxidation product factor. The behaviour of the factors as time series were similar, when considering the differences in the calculation of the factor time series matrix in different SDRTs. For example, the EFA and PCA factors were nearly identical, and the differences in the interpretation lays more in the definition; principal components are defined as linear combinations of the variables (ions) whereas in EFA the variables are expressed as linear combinations of the acquired factors. From the particle phase data, ~~PMF and~~ NMF ~~was~~ere able to separate ~~four~~4 factors, ~~whereas PMF separated two~~. PAM was not able to find more than two separate clusters, most likely due to the high degree of fragmentation in the data and the constrain of PAM to assign one ion to only one cluster, as discussed in Sect. 4.2.2. EFA and PCA had computational constraints due to the small data size acquired from

the AMS and could not to be applied. [In addition, PMF also faced assumedly computational issues with the small particle phase data set, thus not being able to reasonably separate the HOA factor.](#)

The difference, which might be an advantage or disadvantage depending on the application, of PCA and EFA over PMF and NMF is their use of the correlations of the variables instead of the raw data. When using the raw data, ions with high concentration may dominate and hide interesting behaviour occurring in the lower-concentration ions and instead classify those as insignificant background. When using correlations, the concentrations of the ions do not affect the created factorization until the factor time series are calculated, and in principle, variables with different units can be factorized simultaneously. On the other hand, it may diminish some of the more minor and subtle changes. As NMF and PMF do not rely on the correlations, they are more sensitive to the smaller changes taking place in the data. The disadvantage of signals with high intensity dominating in the analysis can be tackled in PMF by choosing an appropriate error matrix that weights the ion signals. Selection of the error matrix can also be crucial when interpreting the PMF output, as a sub-optimal choice may hide the identification of important properties of the data.

The gas-phase data results from PAM agreed moderately with those from EFA and PCA, when taking into account the ability of PAM to assign one ion to only one cluster instead of multiple ones. When comparing the performance of the SDRTs to the bootstrap-type resampled data, we noted that the factorizations from EFA, PCA and PAM were more robust compared to the PMF and NMF results. Results from PMF with different error schemes were similar, but the static error provided more robust solutions when applied to the bootstrap-type resamples.

The findings by Koss et al., (2020) proposed that HCA can be used to quickly identify major patterns in mass spectra data sets, which is in agreement also with our results from PAM. Our findings for PMF partly differ, as they suggest that PMF is not able to sort chemical species into clear generations by their oxidation state. In our study, we found 3 factors (factors 1, 2 and 5, see Table 1), which can be interpreted as representatives for different oxidation states. However, they can also present reactions taking place with different reaction kinetics (faster and slower reactions), as discussed in the results. In addition, Koss et al., (2020) used [only gas-phase data from I-ToF-CIMS and PTR3 with NH₄⁺ as a reagent ion](#), which [are](#) more sensitive to later generation oxidation products compared to [the PTR-PTR-MS](#) which we have used here. We have also used slightly different error types for PMF, which we showed to have a significant impact on the resolved factors, especially if the data size is small. Our results from PMF and NMF agreed reasonably well, even though NMF does not use an error matrix as an input, and it solves the bilinear equation with a different algorithm, indicating our PMF factorization is reasonable and correctly interpreted.

From a mathematical point of view, the selection of the most useful SDRT does not depend on the instrument used to measure the data, nor the extent of fragmentation taking place in the instrument. Only PAM is an exception here, as clustering techniques in general do not assign variables to multiple clusters (i.e., “between” clusters), whereas all the other presented SDRTs have the ability to share an ion between multiple factors. Similarly, if a large number of isomers is to be expected, NMF or PMF may be preferable over EFA or PCA, as the latter two try to maximize the contribution of an ion to a single factor. Ultimately, however, the most useful choice of SDRT also depends on what kind of chemical processes are expected and measured, as the

splitting of ions to multiple factors generally makes the interpretation of the factors more difficult, especially if the prevalence of possible isomerization is not known. Splitting of ions to multiple factors is also an important topic to discuss in source apportionment analysis, where an ion with specific m/z may emerge from various sources or source processes. However, it is a very subtle choice between possibly dismissing unexpected feature discovered by SDRT and using prior knowledge to validate the factorization results. Therefore, applying more than one SDRT may protect the user for determining surprising results “unphysical”, and thus erroneous, but it also gives more robust outcome for the research when the results from different techniques agree.

6 Author’s contribution

SI and SM designed the comparison study; EK and AV designed and organized the measurements/provided data; SI, LH, SM, and AB participated in data analysis and/or interpretation; SI wrote the manuscript; SM, AB, SS, LH, and AV edited the manuscript.

7 Acknowledgments

This work was supported by The Academy of Finland Centre of Excellence (grant no. [307331](#)), [The Academy of Finland Competitive funding to strengthen university research profiles \(PROFI\) for the University of Eastern Finland \(grant no. 325022307331\)](#) and The Nessling foundation. Data collection for this study has been partly funded from the European Union’s 10 Horizon 2020 research and innovation programme through the EUROCHAMP-2020 Infrastructure Activity (grant no. 730997).

Ville Leinonen is thanked for the help he provided with the R software and constructing the error matrices for PMF for the gas phase measurements.

20 8 Competing interest

The authors declare no competing interests.

9 Data availability

The data ~~will be opened~~[can be found](#) in [the EUROCHAMP database](#) ~~upon publication~~ <https://data.eurochamp.org/>

10 References

- Allan, J. D., Jimenez, J. L., Williams, P. I., Alfarra, M. R., Bower, K. N., Jayne, J. T., Coe, H., and Worsnop, D. R.: Quantitative sampling using an Aerodyne aerosol mass spectrometer: 1. Techniques of data interpretation and error analysis, *J Geophys Res-Atmos*, 108, 10.1029/2003jd001607, 2003.
- 5 Brunet, J. P., Tamayo, P., Golub, T. R., and Mesirov, J. P.: Metagenes and molecular pattern discovery using matrix factorization, *Proceedings of the National Academy of Sciences of the United States of America*, 101, 4164-4169, 10.1073/pnas.0308531101, 2004.
- Cattell, R. B.: The scree test for the number of factors. *Multivariate behavioral research*, *Multivar Behav Res*, 1, 245-276, 1966.
- 10 Chakraborty, A., Bhattu, D., Gupta, T., Tripathi, S. N., and Canagaratna, M. R.: Real-time measurements of ambient aerosols in a polluted Indian city: Sources, characteristics, and processing of organic aerosols during foggy and nonfoggy periods, *J Geophys Res-Atmos*, 120, 9006-9019, 10.1002/2015JD023419, 2015.
- Chen, H. Y., Teng, Y. G., Wang, J. S., Song, L. T., and Zuo, R.: Source apportionment of sediment PAHs in the Pearl River Delta region (China) using nonnegative matrix factorization analysis with effective weighted variance solution, *Sci Total Environ*, 444, 401-408, 10.1016/j.scitotenv.2012.11.108, 2013.
- 15 Cleveland, W. S., Grosse, E., and W. M. S.: Local regression models. Chapter 8 of *Statistical Models in S*, edited by: Chambers, J. M., and Hastie, T. J., Wadsworth & Brooks/Cole, 1992.
- Comrey, A. L.: *A First Course in Factor Analysis*, Academic Press, New York, 1973.
- Corbin, J. C., Lohmann, U., Sierau, B., Keller, A., Burtscher, H., and Mensah, A. A.: Black carbon surface oxidation and organic composition of beech-wood soot aerosols, *Atmos. Chem. Phys.*, 15, 11885-11907, 10.5194/acp-15-11885-2015, 2015.
- 20 de Winter, J. C. F., and Dodou, D.: Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size, *J Appl Stat*, 39, 695-710, 10.1080/02664763.2011.610445, 2012.
- Devarajan, K.: Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology, *Plos Comput Biol*, 4, 10.1371/journal.pcbi.1000029, 2008.
- Efron, B., and Tibshirani, R.: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, *Statistical Science*, 1, 54-75, 10.1214/ss/1177013815, 1986.
- 25 Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J.: Evaluating the use of exploratory factor analysis in psychological research, *Psychol Methods*, 4, 272-299, 10.1037/1082-989x.4.3.272, 1999.
- Field, A.: *Discovering Statistics using SPSS*, 4 ed., SAGE, London, 2013.
- Fuller, E. L., and Hemmerle, J. W. J.: Robustness of the maximum-likelihood estimation procedure in factor analysis, *Psychometrika*, 31, 255-266, 1966.
- 30 Gaujoux, R., and Seoighe, C.: A flexible R package for nonnegative matrix factorization, *Bmc Bioinformatics*, 11, 10.1186/1471-2105-11-367, 2010.
- Ghasemi, A., and Zahediasl, S.: Normality tests for statistical analysis: a guide for non-statisticians, *International Journal of Endocrinology and Metabolism*, 10, 486-489, 10.5812/ijem.3505, 2012.
- 35 Golub, G. H., and Van Loan, C. F.: *Matrix Computations*, 3 ed., The Johns Hopkins University Press, Baltimore, 1996.
- Hao, L. Q., Kortelainen, A., Romakkaniemi, S., Portin, H., Jaatinen, A., Leskinen, A., Komppula, M., Miettinen, P., Sueper, D., Pajunoja, A., Smith, J. N., Lehtinen, K. E. J., Worsnop, D. R., Laaksonen, A., and Virtanen, A.: Atmospheric submicron aerosol composition and particulate organic nitrate formation in a boreal forestland-urban mixed region, *Atmos Chem Phys*, 14, 13483-13495, 10.5194/acp-14-13483-2014, 2014.
- 40 Harman, H. H.: *Modern Factor Analysis*, The University of Chicago Press, 1976.
- Horn, J. L.: A rationale and test for the number of factors in factor analysis, *Psychometrika*, 30, 179-185, 1965.
- Hotelling, H.: Analysis of a complex of statistical variables into principal components, *Journal of educational psychology*, 26, 417-441, 1933.
- Hu, L. T., and Bentler, P. M.: Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification, *Psychol Methods*, 3, 424-453, Doi 10.1037/1082-989x.3.4.424, 1998.
- 45 Huang, S. L., Rahn, K. A., and Arimoto, R.: Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island, *Atmos Environ*, 33, 2169-2185, 10.1016/S1352-2310(98)00324-0, 1999.
- Izquierdo, I., Olea, J., and Abad, F. J.: Exploratory factor analysis in validation studies: uses and recommendations, *Psicothema*, 26, 395-400, 10.7334/psicothema2013.349, 2014.
- 50 Jolliffe, I. T.: *Principal Component Analysis*, 2 ed., Springer Series in Statistics, Springer, 2002.
- Kaiser, H. F.: The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, 23, 187-200, 1958.
- Kaiser, H. F.: The application of electronic computers to factor analysis, *Educational and psychological measurement*, 20, 141-151, 1960.
- Kari, E., Miettinen, P., Yli-Pirila, P., Virtanen, A., and Faiola, C. L.: PTR-ToF-MS product ion distributions and humidity-dependence of biogenic volatile organic compounds, *Int J Mass Spectrom*, 430, 87-97, 10.1016/j.ijms.2018.05.003, 2018.

- Kari, E., Faiola, C. L., Isokääntä, S., Miettinen, P., Yli-Pirilä, P., Buchholz, A., Kivimäenpää, M., Mikkonen, S., Holopainen, J. K., and Virtanen, A.: Time-resolved characterization of biotic stress emissions from Scots pines being fed upon by pine weevil by means of PTR-ToF-MS, *Boreal Environ Res*, 24, 25-49, 2019a.
- 5 Kari, E., Hao, L., Ylisirniö, A., Buchholz, A., Leskinen, A., Yli-Pirilä, P., Nuutinen, I., Kuuspalo, K., Jokiniemi, J., Faiola, C., Schobesberger, S., and Virtanen, A.: Potential dual effect of anthropogenic emissions on the formation of biogenic secondary organic aerosol (BSOA) *Atmos. Chem. Phys.*, 19, 15651-15671, 10.5194/acp-19-15651-2019, 2019b.
- Kassambara, A., and Mundt, F.: *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2017.
- Kaufman, L., and Rousseeuw, P. J.: *Finding groups in data : an introduction to cluster analysis*, Wiley series in probability and mathematical statistics, Applied probability and statistics, Wiley, New York, 342 pp., 1990.
- 10 Kim, H. J.: *Common Factor Analysis Versus Principal Component Analysis: Choice for Symptom Cluster Research*, *Asian Nurs Res*, 2, 17-24, 10.1016/S1976-1317(08)60025-0, 2008.
- Kim, P. M., and Tidor, B.: Subsystem identification through dimensionality reduction of large-scale gene expression data, *Genome Res*, 13, 1706-1718, 10.1101/gr.903503, 2003.
- Korkmaz, S., Goksuluk, D., and Zararsiz, G.: MVN: An R Package for Assessing Multivariate Normality, *R J*, 6, 151-162, 2014.
- 15 Kortelainen, A., Joutsensaari, J., Hao, L., Leskinen, J., Tiitta, P., Jaatinen, A., Miettinen, P., Sippula, O., Torvela, T., Tissari, J., Jokiniemi, J., Worsnop, D. R., Smith, J. N., Laaksonen, A., and Virtanen, A.: Real-Time Chemical Composition Analysis of Particulate Emissions from Woodchip Combustion, *Energy Fuel*, 29, 1143-1150, 10.1021/ef5019548, 2015.
- Koss, A. R., Canagaratna, M. R., Zaytsev, A., Krechmer, J. E., Breitenlechner, M., Nihill, K. J., Lim, C. Y., Rowe, J. C., Roscioli, J. R., Keutsch, F. N., and Kroll, J. H.: Dimensionality-reduction techniques for complex mass spectrometric datasets: application to laboratory atmospheric organic oxidation experiments, *Atmos Chem Phys*, 20, 1021-1041, 10.5194/acp-20-1021-2020, 2020.
- 20 Lee, D. D., and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, *Nature*, 401, 788-791, 10.1038/44565, 1999.
- Lee, D. D., and Seung, H. S.: Algorithms for non-negative matrix factorization, *Advances in Neural and Information Processing Systems*, 13, 556-562, 2001.
- Leskinen, A., Yli-Pirilä, P., Kuuspalo, K., Sippula, O., Jalava, P., Hirvonen, M. R., Jokiniemi, J., Virtanen, A., Komppula, M., and Lehtinen, K. E. J.: Characterization and testing of a new environmental chamber, *Atmos Meas Tech*, 8, 2267-2278, 10.5194/amt-8-2267-2015, 2015.
- 25 Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K.: *cluster: Cluster Analysis Basics and Extensions*, 2018.
- Malley, C. S., Braban, C. F., and Heal, M. R.: The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification, *Atmos Res*, 138, 30-40, 10.1016/j.atmosres.2013.10.019, 2014.
- 30 Massoli, P., Stark, H., Canagaratna, M. R., Krechmer, J. E., Xu, L., Ng, N. L., Mauldin, R. L., Yan, C., Kimmel, J., Misztal, P. K., Jimenez, J. L., Jayne, J. T., and Worsnop, D. R.: Ambient Measurements of Highly Oxidized Gas-Phase Molecules during the Southern Oxidant and Aerosol Study (SOAS) 2013, *ACS Earth Space Chem*, 2, 653-672, 10.1021/acsearthspacechem.8b00028, 2018.
- Morrison, D.: *Multivariate Statistical Methods*, 4 ed., Thomson/Brooks/Cole, Belmont, CA, 2005.
- National Research Council: *Rethinking the Ozone Problem in Urban and Regional Air Pollution*, The National Academies Press, Washington, DC, 524 pp., 1991.
- 35 Onasch, T. B., Trimborn, A., Fortner, E. C., Jayne, J. T., Kok, G. L., Williams, L. R., Davidovits, P., and Worsnop, D. R.: Soot Particle Aerosol Mass Spectrometer: Development, Validation, and Initial Application, *Aerosol Sci Tech*, 46, 804-817, 10.1080/02786826.2012.663948, 2012.
- Osborne, J.: *Best Practices in Exploratory Factor Analysis*, CreateSpace Independent Publishing Platform, 2014.
- 40 Paatero, P., and Tapper, U.: Analysis of Different Modes of Factor-Analysis as Least-Squares Fit Problems, *Chemometr Intell Lab*, 18, 183-194, 10.1016/0169-7439(93)80055-M, 1993.
- Paatero, P., and Tapper, U.: Positive Matrix Factorization - a Nonnegative Factor Model with Optimal Utilization of Error-Estimates of Data Values, *Environmetrics*, 5, 111-126, 10.1002/env.3170050203, 1994.
- Paatero, P.: Least squares formulation of robust non-negative factor analysis, *Chemometr Intell Lab*, 37, 23-35, 10.1016/S0169-7439(96)00044-5, 1997.
- 45 Paatero, P., Hopke, P. K., Song, X. H., and Ramadan, Z.: Understanding and controlling rotations in factor analytic models, *Chemometr Intell Lab*, 60, 253-264, 10.1016/S0169-7439(01)00200-3, 2002.
- Paatero, P., and Hopke, P. K.: Discarding or downweighting high-noise variables in factor analytic models, *Anal Chim Acta*, 490, 277-289, 10.1016/S0003-2670(02)01643-4, 2003.
- 50 Paatero, P.: Interactive comment on "Source characterization of Highly Oxidized Multifunctional Compounds in a Boreal Forest Environment using Positive Matrix Factorization" by Chao Yan et al., *Atmospheric Chemistry and Physics Discussion*, 2016.
- Pandit, S., and Gupta, S.: *International Journal of Research in Computer Science*, 2, 29-31, 10.7815/ijorcs.21.2011.011, 2011.
- Pearson, K.: On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 559-572, 1901.
- 55 Pekey, H., Bakoglu, M., and Pekey, B.: Sources of heavy metals in the Western Bay of Izmit surface sediments, *Int J Environ an Ch*, 85, 1025-1036, 10.1080/03067310500194953, 2005.

- R Core Team: R: A language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria, 2019.
- Raskin, R., and Terry, H.: A Principal-Components Analysis of the Narcissistic Personality-Inventory and Further Evidence of Its Construct-Validity, *J Pers Soc Psychol*, 54, 890-902, 10.1037/0022-3514.54.5.890, 1988.
- Rencher, A., and Christensen, W.: *Methods of Multivariate Analysis*, 3 ed., Wiley Series in Probability and Statistics, Wiley, 796 pp., 2012.
- 5 Revelle, W.: *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois, 2018.
- Rosati, B., Teiwes, R., Kristensen, K., Bossi, R., Skov, H., Glasius, M., Pedersen, H. B., and Bilde, M.: Factor analysis of chemical ionization experiments: Numerical simulations and an experimental case study of the ozonolysis of alpha-pinene using a PTR-ToF-MS, *Atmos Environ*, 199, 15-31, 10.1016/j.atmosenv.2018.11.012, 2019.
- Ruscio, J., and Roche, B.: Determining the Number of Factors to Retain in an Exploratory Factor Analysis Using Comparison Data of Known Factorial Structure, *Psychol Assessment*, 24, 282-292, 10.1037/a0025697, 2012.
- 10 Sánchez-López, J. A., Zimmermann, R., and Yeretzyan, C.: Insight into the time-resolved extraction of aroma compounds during espresso coffee preparation: online monitoring by PTR-ToF-MS., *Anal Chem*, 86, 11696-11704, 2014.
- Schwarz, G.: Estimating the Dimension of a Model. *Annals of Statistics*, 6, 461-464, doi:10.1214/aos/1176344136, 1978.
- Seinfeld, J. H., and Pandis, S. N.: *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 3 ed., John Wiley & Sons, Hoboken, New Jersey, 2016.
- 15 Sofowote, U. M., McCarty, B. E., and Marvin, C. H.: Source apportionment of PAH in Hamilton Harbour suspended sediments: Comparison of two factor analysis methods, *Environ Sci Technol*, 42, 6007-6014, 10.1021/es800219z, 2008.
- Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., and Satoto, B. D.: Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster, *IOP Conference Series: Materials Science and Engineering*, 336, 10.1088/1757-899X/336/1/012017, 2018.
- 20 Tabachnick, B. G., and Fidell, L. S.: *Using Multivariate Statistics*, 6 ed., Pearson, 2014.
- Tibshirani, R., Walther, G., and Hastie, T.: Estimating the number of clusters in a data set via the gap statistic, *J Roy Stat Soc B*, 63, 411-423, Doi 10.1111/1467-9868.00293, 2001.
- Tiitta, P., Leskinen, A., Hao, L., Yli-Pirilä, P., Kortelainen, M., Grigonyte, J., Tissari, J., Lamberg, H., Hartikainen, A., Kuusalo, K., 25 Kortelainen, A. M., Virtanen, A., Lehtinen, K. E. J., Komppula, M., Pieber, S., Prévôt, A. S. H., Onasch, T. B., Worsnop, D. R., Czech, H., Zimmermann, R., Jokiniemi, J., and Sippula, O.: Transformation of logwood combustion emissions in a smog chamber: formation of secondary organic aerosol and changes in the primary organic aerosol upon daytime and nighttime aging, *Atmos. Chem. Phys.*, 16, 13251-13269, 10.5194/acp-16-13251-2016, 2016.
- Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, *Atmos Chem Phys*, 9, 2891-2918, 10.5194/acp-9-2891-2009, 2009.
- 30 Vigneau, E.: *ClustVarLV: Clustering of Variables Around Latent Variables*. 2016.
- Wold, S., Esbensen, K., and Geladi, P.: Principal Component Analysis, *Chemometr Intell Lab*, 2, 37-52, 10.1016/0169-7439(87)80084-9, 1987.
- Wu, J.: *Advances in K-means Clustering: A Data Mining Thinking*, Springer Theses, Springer, 2012.
- 35 Wyche, K. P., Monks, P. S., Smallbone, K. L., Hamilton, J. F., Alfarra, M. R., Rickard, A. R., McFiggans, G. B., Jenkin, M. E., Bloss, W. J., Ryan, A. C., Hewitt, C. N., and MacKenzie, A. R.: Mapping gas-phase organic reactivity and concomitant secondary organic aerosol formation: chemometric dimension reduction techniques for the deconvolution of complex atmospheric data sets, *Atmos Chem Phys*, 15, 8077-8100, 10.5194/acp-15-8077-2015, 2015.
- 40 Yan, C., Nie, W., Äijälä, M., Rissanen, M. P., Canagaratna, M. R., Massoli, P., Junninen, H., Jokinen, T., Sarnela, N., Hame, S. A. K., Schobesberger, S., Canonaco, F., Yao, L., Prevot, A. S. H., Petaja, T., Kulmala, M., Sipilä, M., Worsnop, D. R., and Ehn, M.: Source characterization of highly oxidized multifunctional compounds in a boreal forest environment using positive matrix factorization, *Atmos Chem Phys*, 16, 12715-12731, 10.5194/acp-16-12715-2016, 2016.
- Yan, M., Yang, X., Hang, W., and Xia, Y.: Determining the number of factors for non-negative matrix and its application in source apportionment of air pollution in Singapore, *Stoch Env Res Risk A*, 33, 1175-1186, 10.1007/s00477-019-01677-z, 2019.
- 45 Zhang, Q., Alfarra, M. R., Worsnop, D. R., Allan, J. D., Coe, H., Canagaratna, M. R., and Jimenez, J. L.: Deconvolution and quantification of hydrocarbon-like and oxygenated organic aerosols based on aerosol mass spectrometry, *Environ Sci Technol*, 39, 4938-4952, 10.1021/es0485681, 2005.
- Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Ulbrich, I. M., Ng, N. L., Worsnop, D. R., and Sun, Y. L.: Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry: a review, *Anal Bioanal Chem*, 401, 3045-3067, 10.1007/s00216-011-5355-y, 2011.
- 50 Äijälä, M., Heikkinen, L., Frohlich, R., Canonaco, F., Prevot, A. S. H., Junninen, H., Petaja, T., Kulmala, M., Worsnop, D., and Ehn, M.: Resolving anthropogenic aerosol pollution types - deconvolution and exploratory classification of pollution events, *Atmos Chem Phys*, 17, 3165-3197, 10.5194/acp-17-3165-2017, 2017.

5

10

Tables

Table 1 Summary of the results for gas phase composition data. Best solution refers to the number of factors/clusters/components. The m/z refers to the mass with the H⁺.

Type of analysis	EFA	EVD- PCA	PAM	NMF	PMF (static error)	PMF (signal following error)	Example compounds
best solution	4	4	4	5	5	5	
rotation if used	Oblimin	Oblimin	-	-	-	-	
precursor (α-pinene)	3	3	3	3	3	3	α -pinene, C ₇ H ₁₀ , Toluene
early products	2	2	2	2	2	2	MVK, furan, acetaldehyde
later products/slowly forming products	1	1	1	1	1	1	C ₂ H ₂ O, C ₂ H ₈ O, CH ₃ O ₂ , MEK
intermediate products	-	-	-	5	5	5	Nopinone, m/z 157.08
precursor (car exhaust) or background	4	4	4	4	4	4	C ₄ H ₈ O ₂ , m/z 167.06, Dimethylbenzene

15

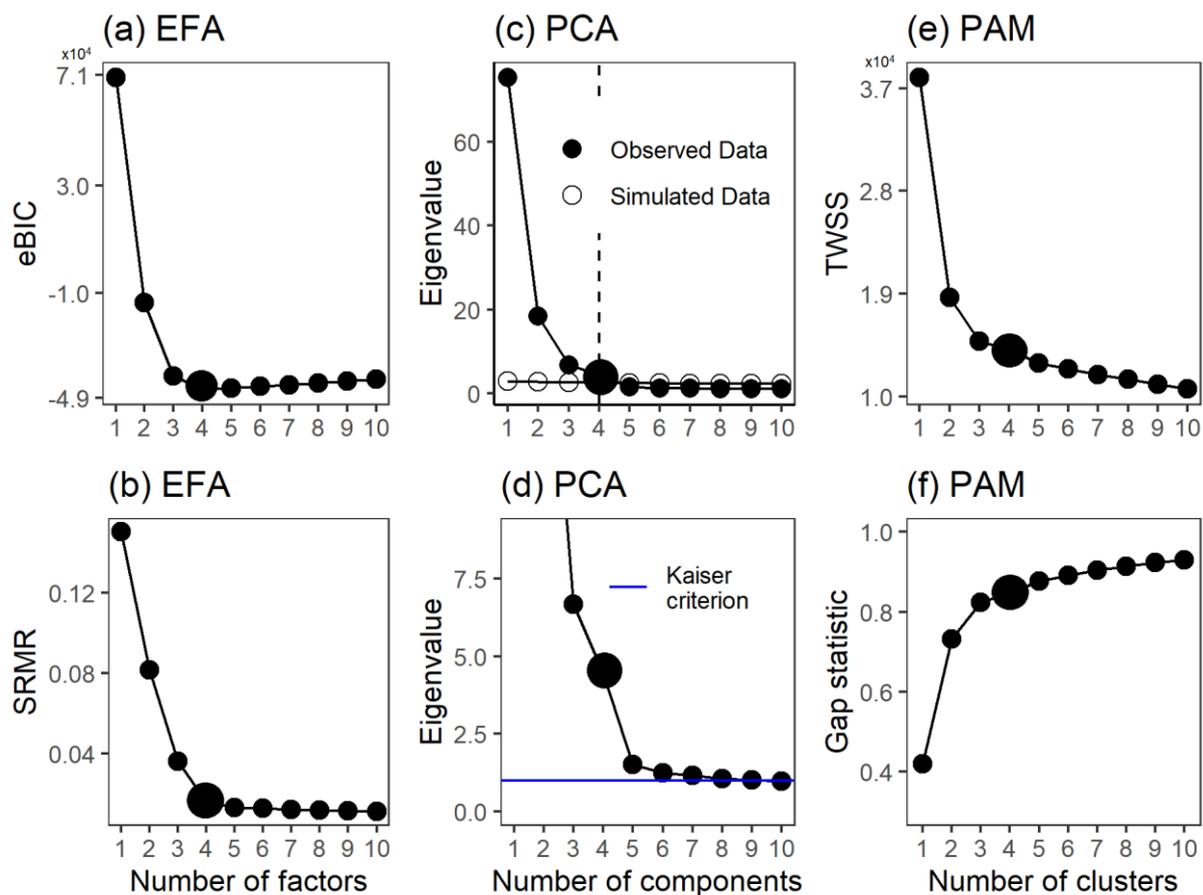
Table 2 Summary of the results for particle phase composition data. Best solution refers to the number of factors/clusters.

<u>Type of analysis</u>	<u>NMF</u>	<u>PMF</u> <u>(static error)</u>	<u>PMF (standard</u> <u>AMS error)</u>	<u>Example</u> <u>compounds</u>
<u>best solution</u>	<u>4</u>	<u>2</u>	<u>2</u>	
<u>primary OA (HOA)</u>	<u>1</u>	-	-	<u>m/z 12, 57, 59</u>
<u>mixed LVOOA</u>	<u>3</u>	-	-	<u>m/z 44</u>
<u>αP-SOA-SVOOA</u>	<u>2</u>	-	-	<u>m/z 43</u>
<u>αP-SOA-LVOOA</u>	<u>4</u>	-	-	<u>m/z 44</u>
<u>SVOOA</u>		<u>1</u>	<u>1</u>	<u>m/z 43</u>
<u>LVOOA</u>		<u>2</u>	<u>2</u>	<u>m/z 44</u>

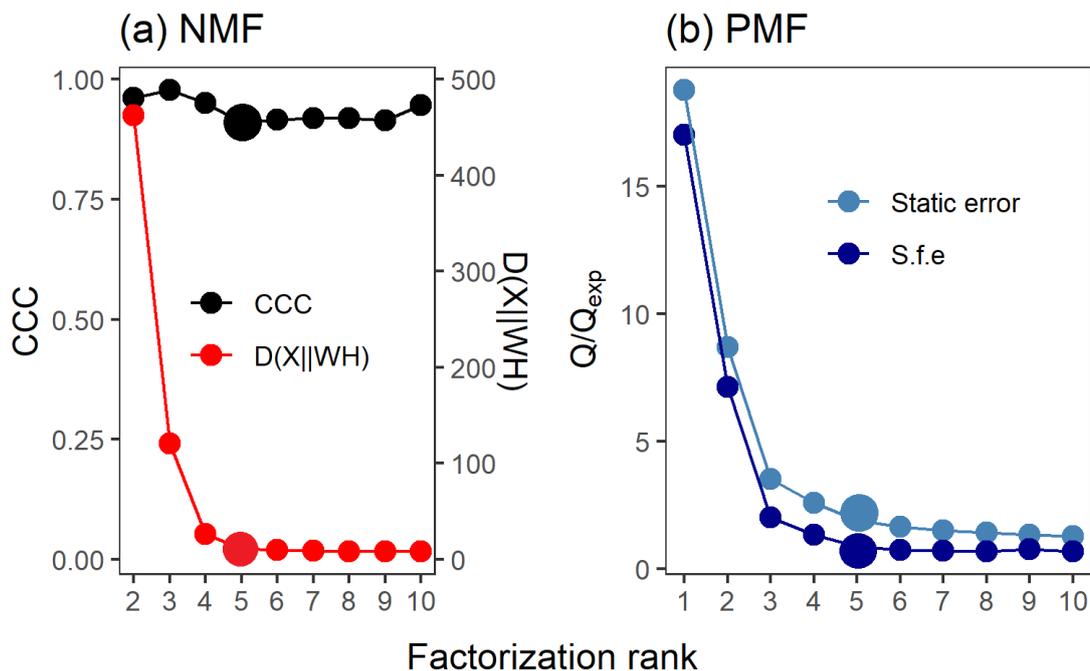
5 **Table 3 The computational time (in seconds) for 2-9 factors for different SDTRs and data types and sizes.**

SDRT	AMS	<u>PTRPTR-</u> <u>MS</u>	<u>PTRPTR-</u> <u>MS*5</u>
EVD-PCA	-	1.93	5.23
SVD-PCA	0.571	0.838	1.59
ml-EFA	-	14.2	16.9
pa-EFA	-	2.96	6.18
PAM	0.672	0.771	1.69
NMF	21.6	39.7	134
PMF (static error)	30.0	101	476
PMF (<u>standard AMS error</u> <u>Poisson or noise error</u>)	31.1	122	543

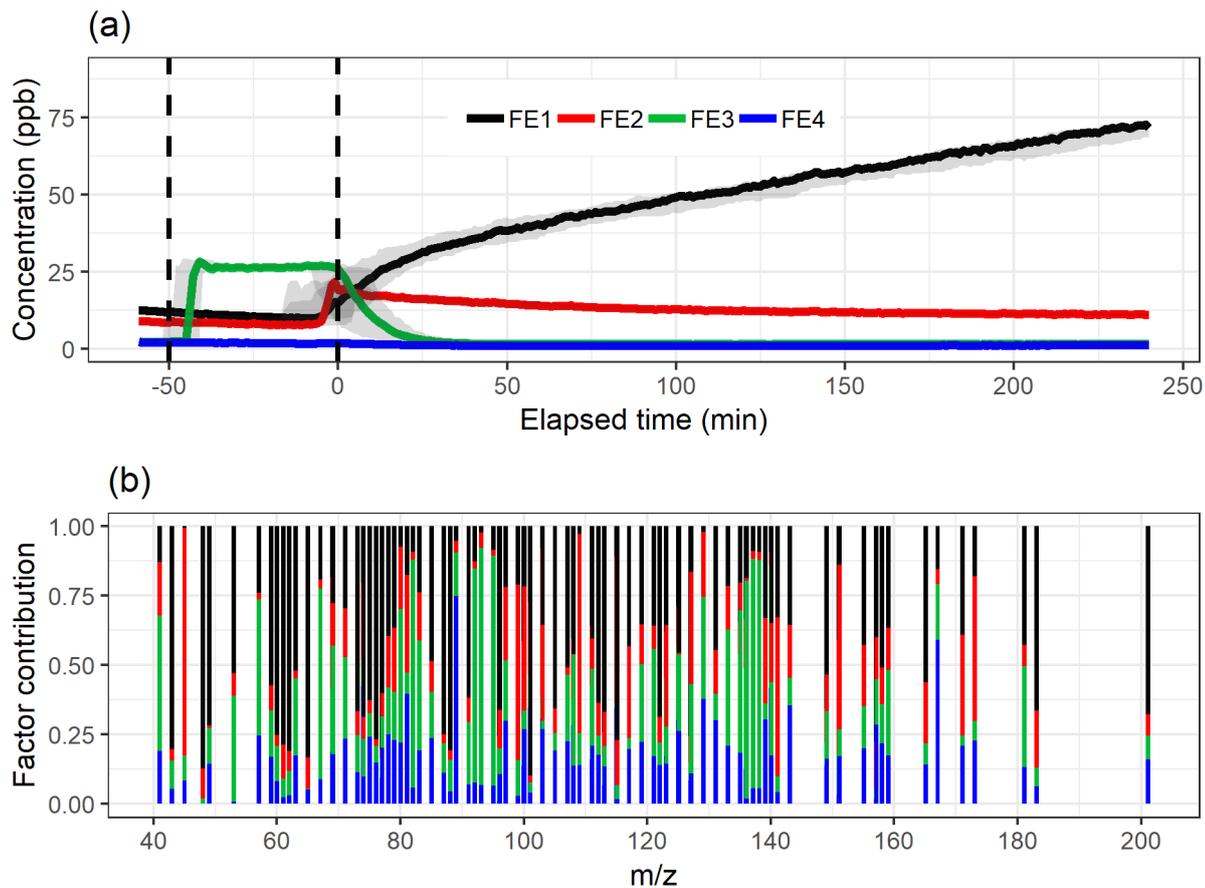
Figures



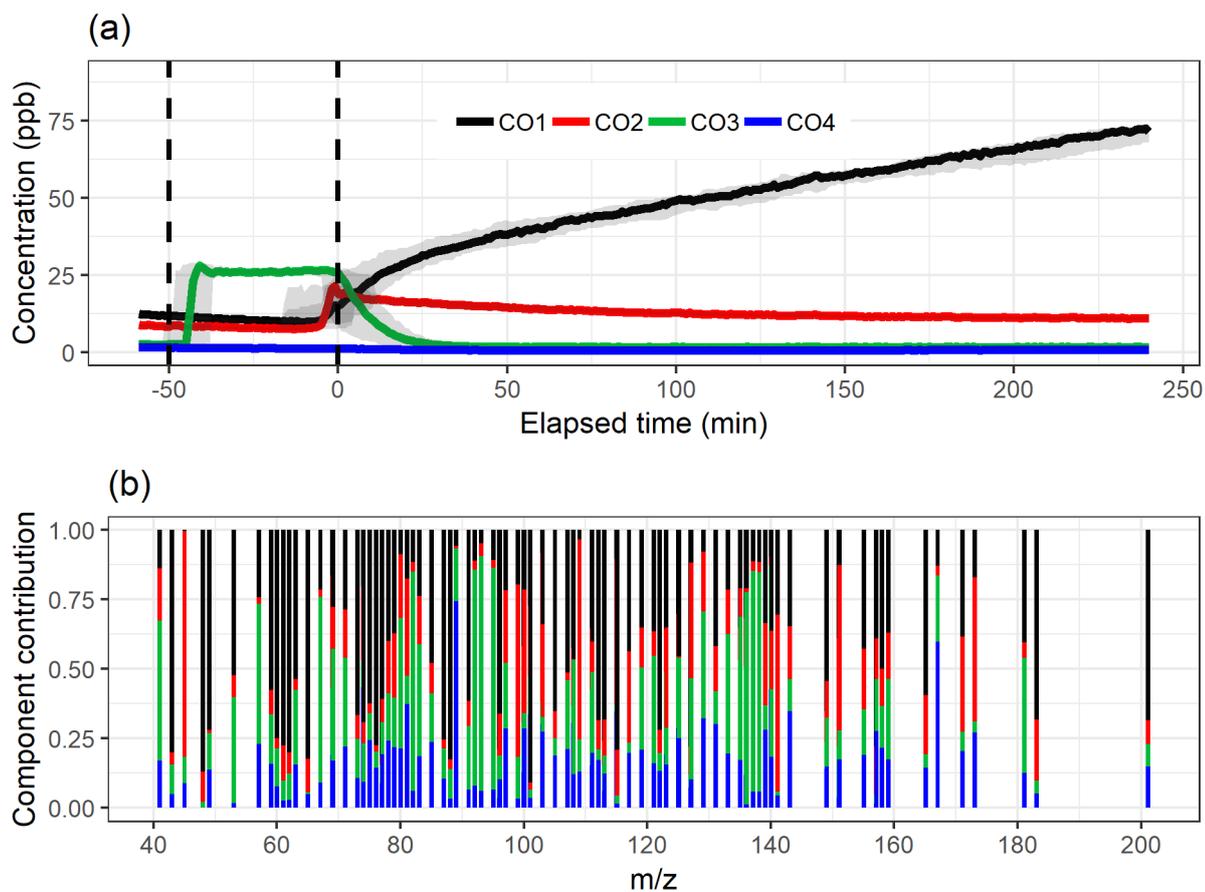
- 5 **Figure 1: Factor number indexes for gas phase data (PTR-MS).** Empirical BIC (a) and SRMR (b) as a function of the number of factors for ml-EFA, Parallel analysis (c) and Kaiser criterion (d) for EVD-PCA and TWSS (e) and Gap statistic (f) for PAM. **Data measured with PTR. Larger points indicate the solution that was selected for more detailed interpretation.**



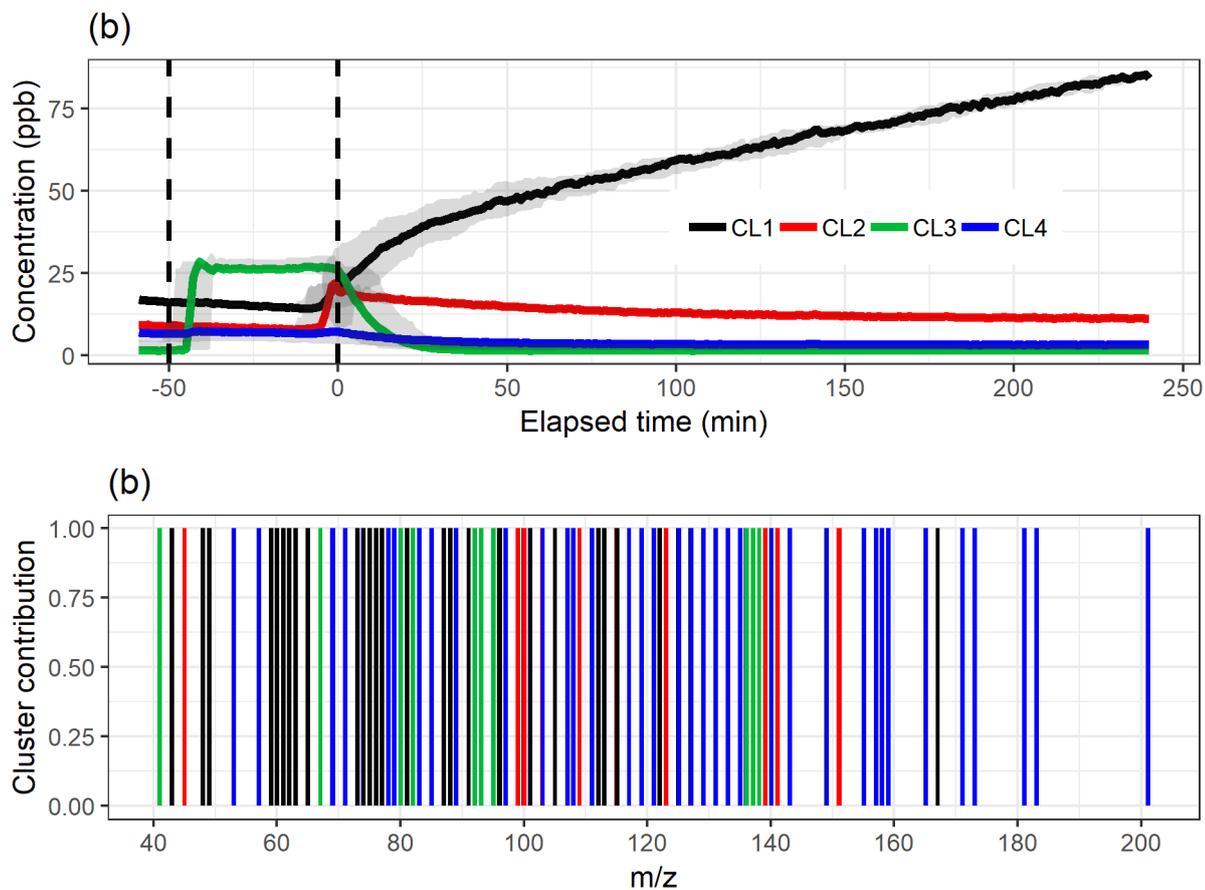
5 **Figure 2: Factor number indexes for gas phase data (PTR-MS). Estimation of the factorization rank for NMF in (a) with CCC and $D(X||WH)$ and for PMF in (b) with Q/Q_{exp} for the two error schemes (static error and signal following error, S.f.e). Data measured with PTR. Larger points indicate the solution that was selected for more detailed interpretation.**



5 **Figure 3: The factor time series (a) and contribution (b) for ml-EFA with Oblimin rotation for the 4-factor solution. Shaded areas in the time series indicate the factor range for the bootstrap resamples, solid lines are for the measured PTR gas phase (PTR-MS) data. The colour code identifying factors is the same in both panels. Factors were identified as later/slowly forming products (FE1), early products (FE2), α -pinene precursor (FE3) and background/car exhaust precursor (FE4).**



5 **Figure 4: The component time series (a) and contribution (b) for EVD-PCA with Oblimin rotation for the 4-component solution. Shaded areas in the time series indicate the component range for the bootstrap resamples, solid lines are for the measured [gas phase PTR \(PTR-MS\)](#) data. The colour code identifying components is the same in both panels. [Components were identified as later/slowly forming products \(CO1\), early products \(CO2\), \$\alpha\$ -pinene precursor \(CO3\) and background/car exhaust precursor \(CO4\).](#)**



5 **Figure 5: The time series (a) of the clusters and the distribution of ion to clusters (b) from PAM with 4-cluster solution. Shaded areas in the time series indicate the cluster range for the bootstrap resamples, solid lines are for the measured PTR-gas phase (PTR-MS) data. The colour code identifying clusters is the same in both panels. Clusters were identified as later/slowly forming products (CL1), early products (CL2), α -pinene precursor (CL3) and background/car exhaust precursor (CL4).**

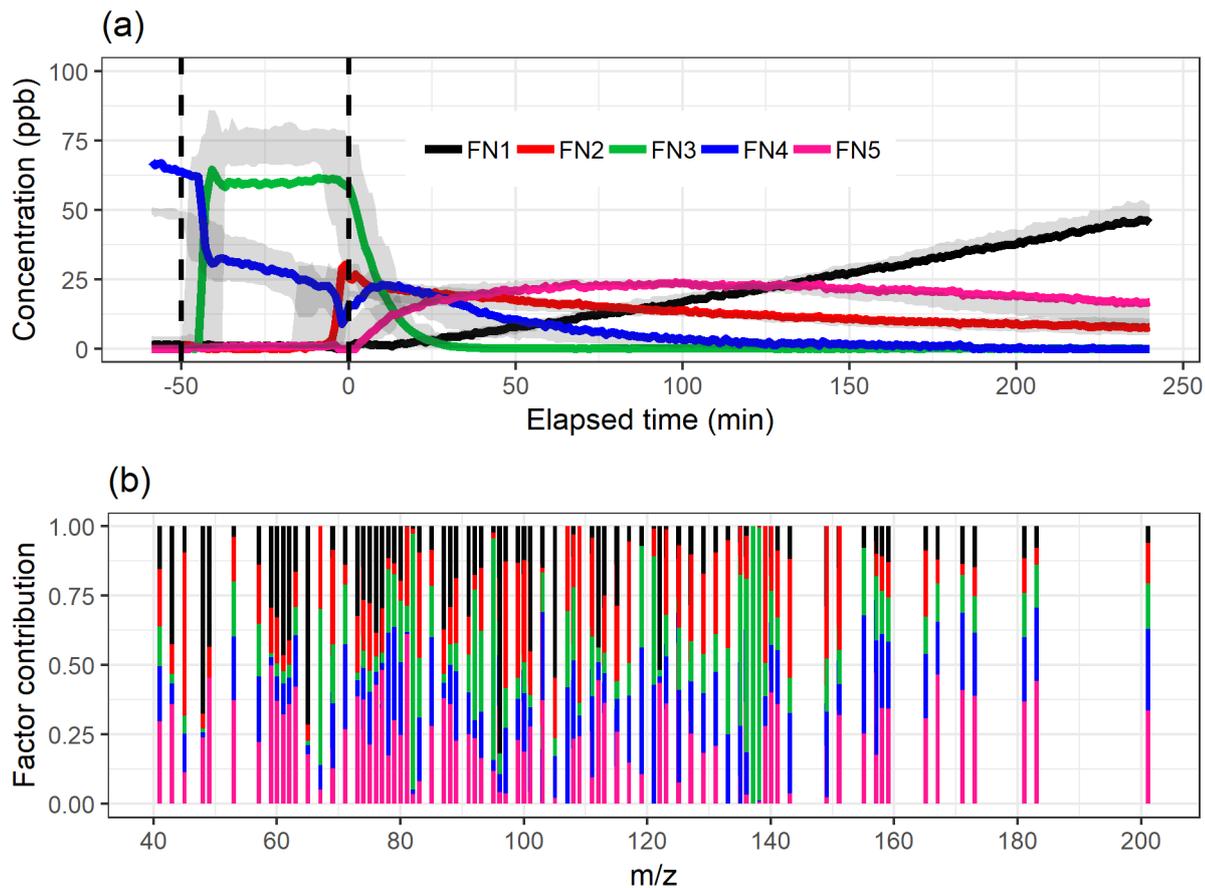
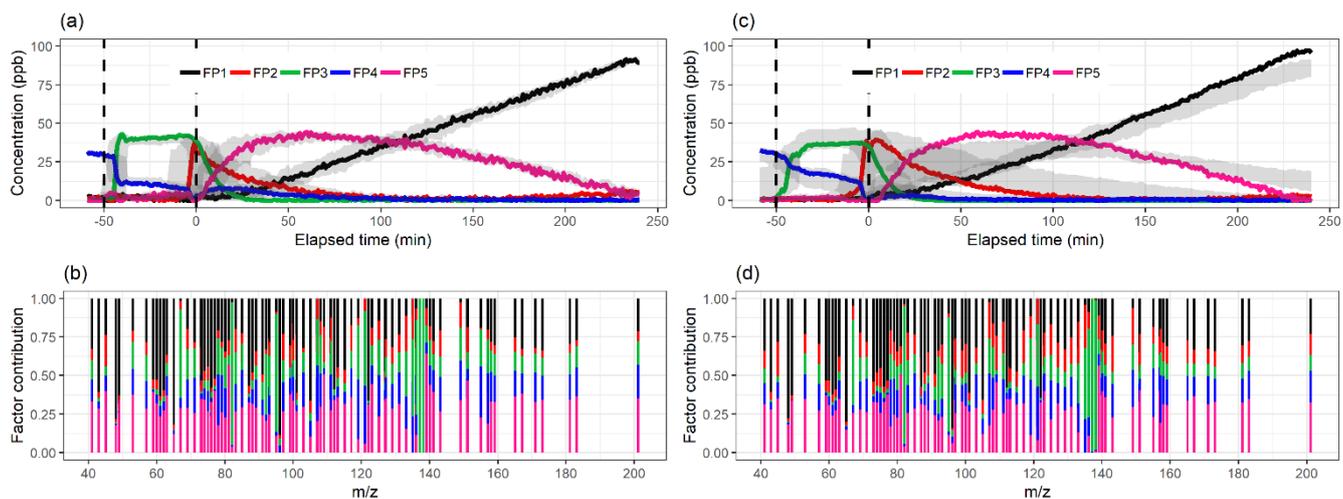


Figure 6: The time series (a) of the factors and the distribution of ion to factor (b) from NMF with 5-factor solution. Shaded areas in the time series indicates the factor range for the bootstrap resamples, solid lines are for the measured PTRgas phase (PTR-MS) data. The colour code identifying clusters is the same in both panels. Factors were identified as later/slowly forming products (FN1), early products (FN2), α -pinene precursor (FN3), background/car exhaust precursor (FN4) and intermediate products (FN5).

5



5 **Figure 7: Factor time series and contribution from PMF with static error (a-b) and signal following error (c-d) for factorization rank 5. Shaded areas in the time series indicates the factor range for the bootstrap resamples, solid lines are for the measured PTRgas phase (PTR-MS) data. The colour code identifying the factors is the same in the top and bottom panels. Factors were identified as later/slowly forming products (FP1), early products (FP2), α -pinene precursor (FP3), background/car exhaust precursor (FP4) and intermediate products (FP5).**

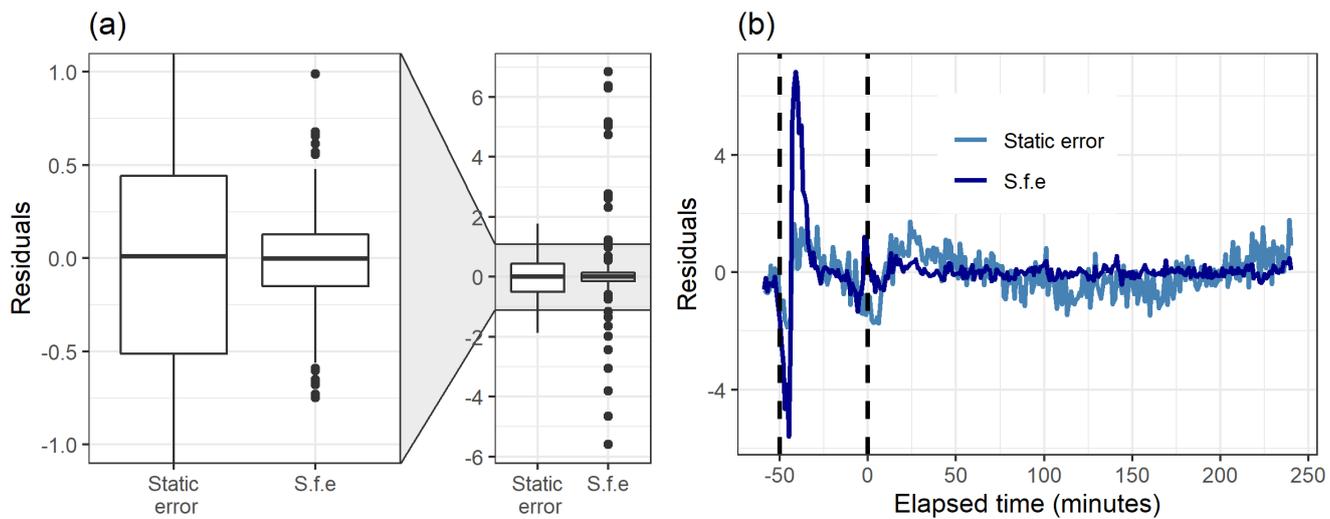
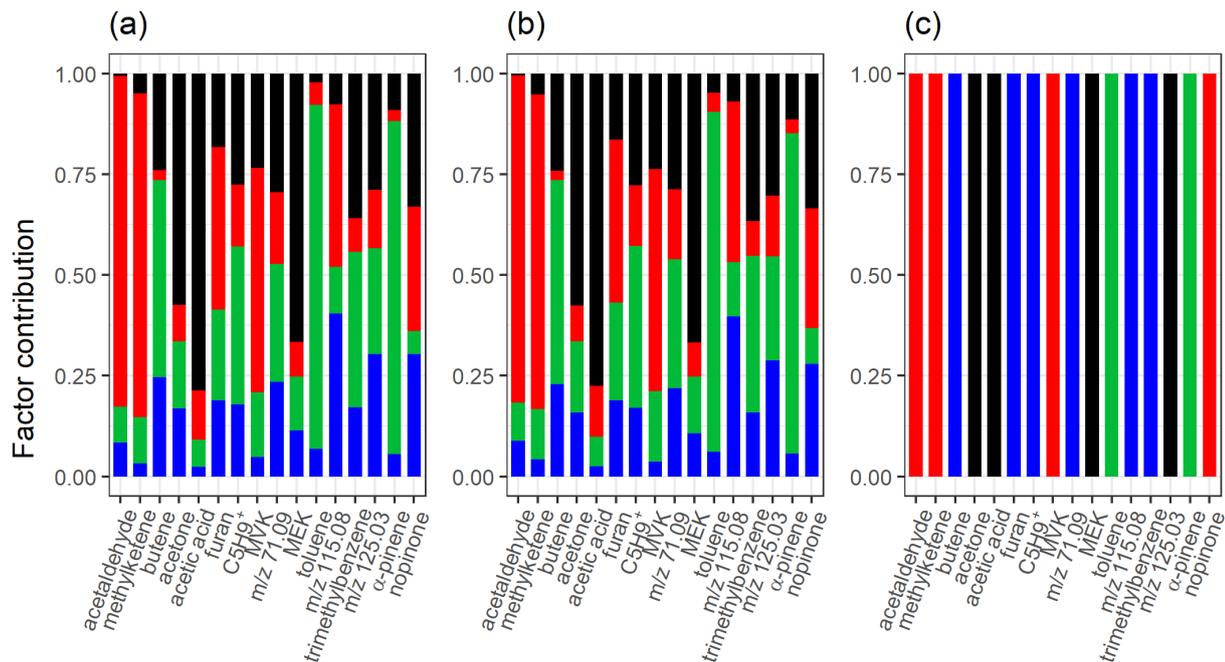
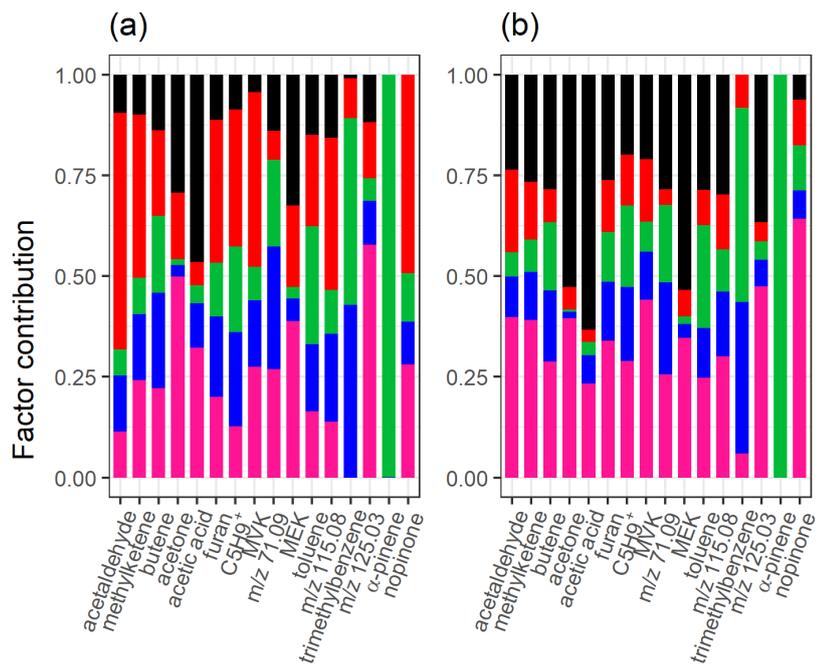


Figure 8: Boxplot (a) and the time series (b) of the residuals (original total signal – reconstructed total signal) with static error and signal following error (S.f.e) with 5 factors from PMF for the measured PTR gas phase (PTR-MS) data.

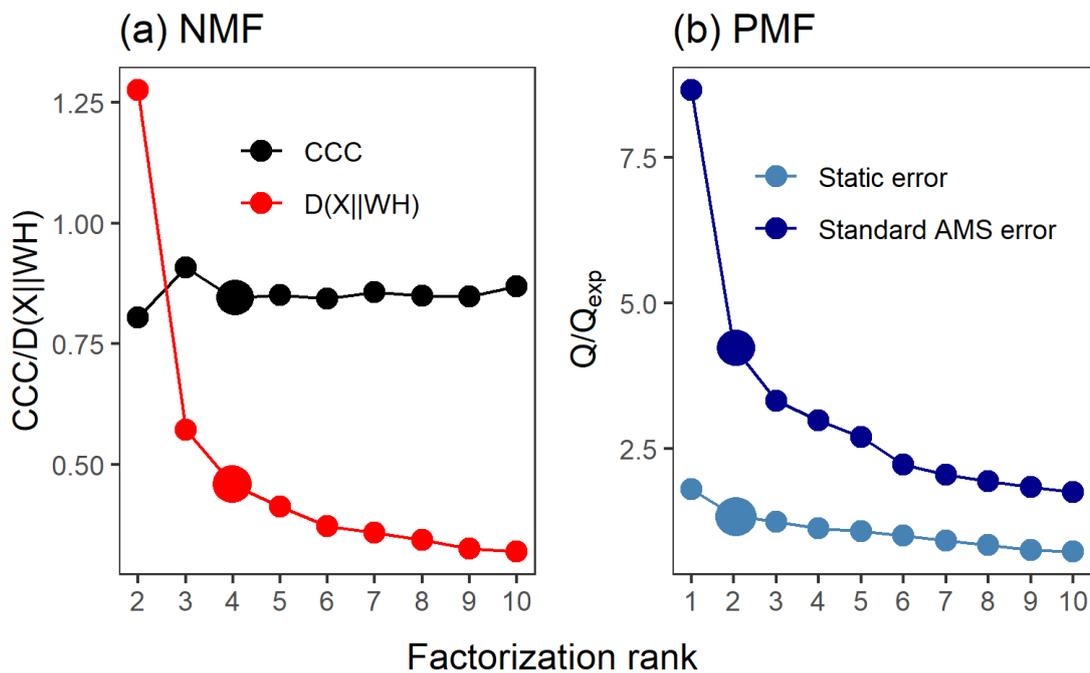


5

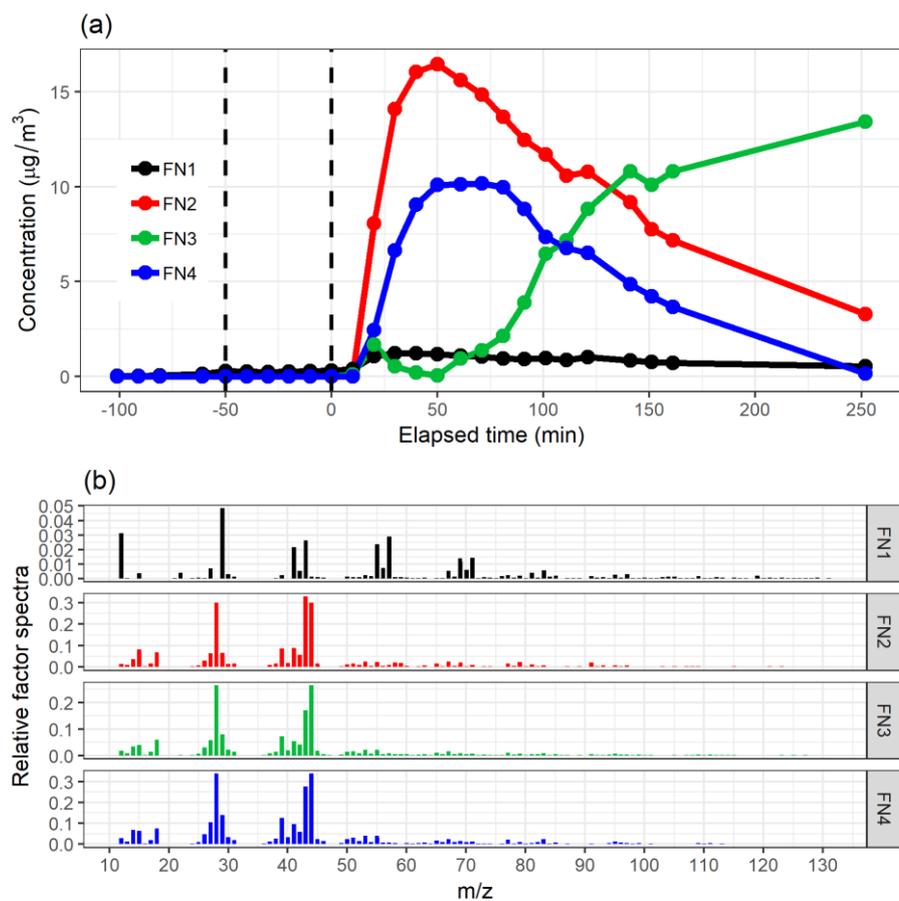
Figure 9: Total factor/component/cluster contribution of selected compounds from a) ml-EFA, b) EVD-PCA and c) PAM for the measured PTRgas phase (PTR-MS) data. Colour code identifying the factors/components/clusters is the same in all panels and were identified as later/slowly forming products (black), early products (red), α -pinene precursor (green) and background/car exhaust precursor (blue).



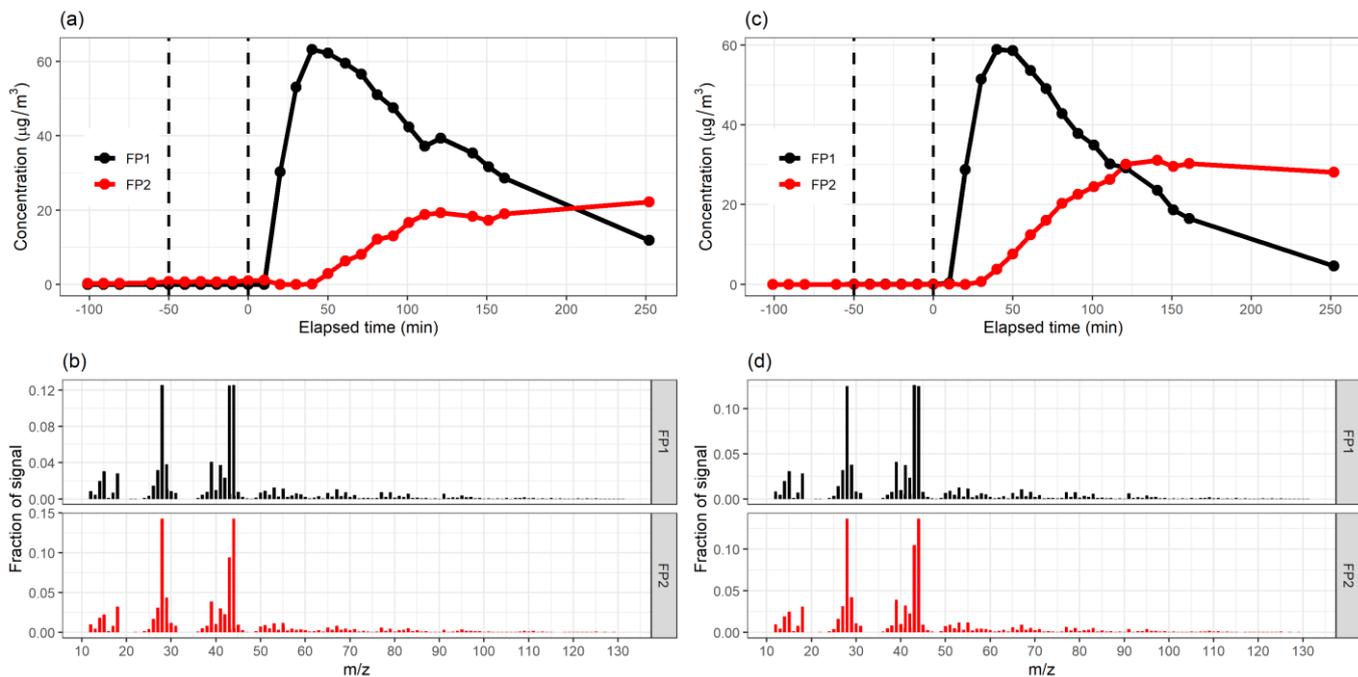
5 **Figure 10: Total factor contribution of selected compounds from a) NMF and b) PMF with static error for the measured PTRgas phase (PTR-MS) data. Colour code identifying the factors is the same in both panels, and were identified as later/slowly forming products (black), early products (red), α -pinene precursor (green), background/car exhaust precursor (blue) and intermediate products (pink).**



5 **Figure 11: [Factor number indexes for particle phase data \(AMS\)](#). Estimation of the factorization rank for NMF in (a) with CCC and [D\(X||WH\)_{RSS}](#) and for PMF in (b) with Q/Q_{exp} for the two error schemes. [Data measured with AMS](#). [Larger points indicate the solution that was selected for more detailed interpretation](#).**

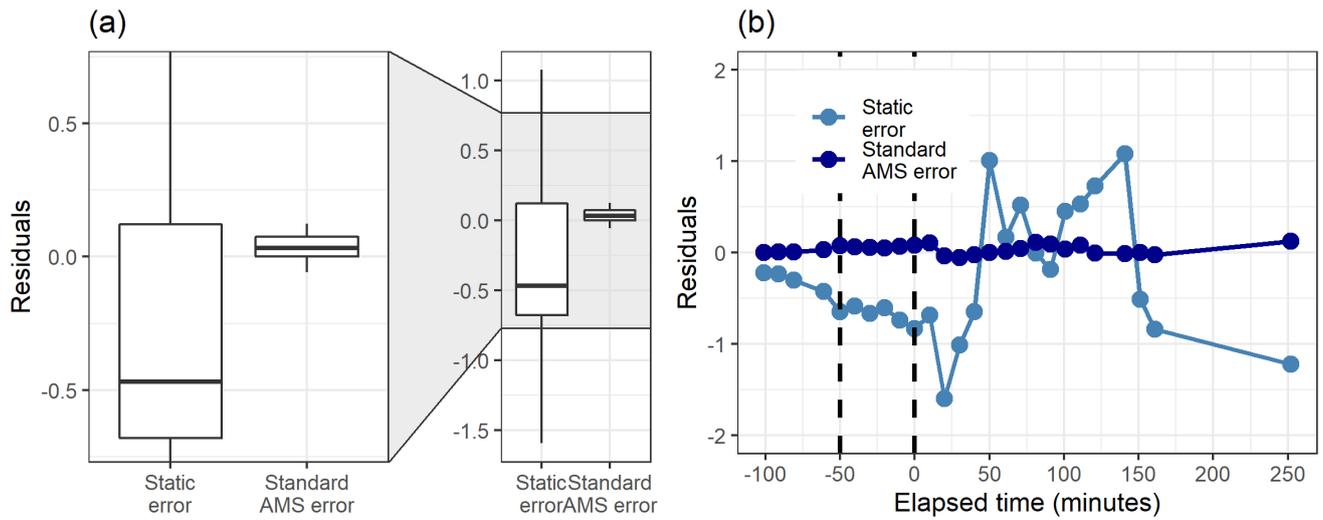


5 **Figure 12: The factor time series (a) and relative factor spectra (b) from NMF with 4 factors for the measured particle phase (AMS) data. The colour code identifying the factors is the same in both panels. Factors were identified as primary OA (FN1), α P-SOA-SVOOA (FN2), mixed LVOOA (FN3) and α P-SOA-LVOOA (FN4).**



5 **Figure 13: Factor time series and contribution from PMF with static error (a-b) and [Poisson-style error standard AMS error](#) (c-d) for factorization rank [24](#) for the measured [particle phase \(AMS\)](#) data. [Factors were identified as SVOOA \(FP1\) and LVOOA \(FP2\)](#)**

10



5 **Figure 14: Boxplot (a) and the time series (b) of the residuals (original total signal – reconstructed total signal) with static error and ~~Poisson style error~~ standard AMS error with 24 factors from PMF for the measured particle phase (AMS) data.**

10

15

20

Appendix A

Table A1 Mathematical symbols and notations used in the equations throughout the paper.

symbol	explanation
\mathbf{X}, X_{ij}	data matrix (n x m), data matrix element
p	number of factors/components/clusters
y_j	variable/ion j (time series vector), column j from \mathbf{X}
c_j	PCA component j
f	EFA factor
λ, λ_{ij}	EFA loading matrix, loading matrix element
\mathbf{S}, \mathbf{R}	observed correlation matrix, implied correlation matrix
\mathbf{G}	factorization matrix (factor time series) PMF (n x p)
\mathbf{F}	factorization matrix (factor spectra/contribution) in PMF (p x m)
$\boldsymbol{\mu}$	PMF error matrix
\mathbf{E}	residual matrix in PMF
\mathbf{W}	factorization matrix (factor time series) in NMF (n x \underline{kp})
\mathbf{H}	factorization matrix (factor spectra/contribution) in NMF (\underline{kp} x m)

Supplementary Information for the manuscript “*Comparison of dimension reduction techniques in the analysis of mass spectrometry data*”

Sini Isokääntä¹, Eetu Kari^{1,a}, Angela Buchholz¹, Liqing Hao¹, Siegfried Schobesberger¹, Annele Virtanen¹, Santtu Mikkonen^{1,2}

5 ¹Department of Applied Physics, University of Eastern Finland, Kuopio, 70210, Finland

²[Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, 70210, Finland](#)

^a[currently at: Neste Oyj, Espoo, Finland](#)

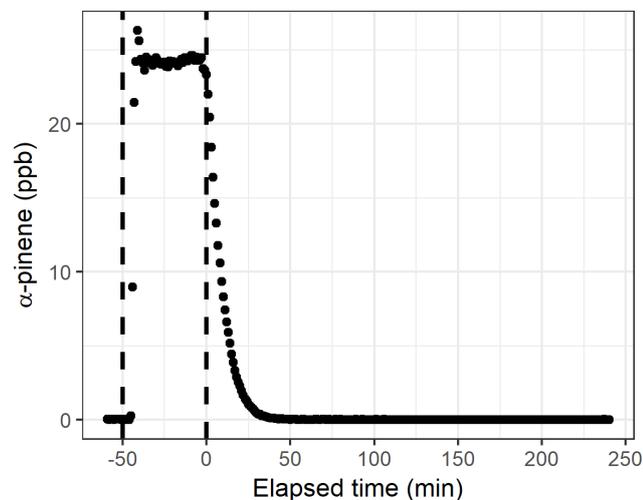
Correspondence to: S. Isokääntä (sini.isokaanta@uef.fi)

10 S1 Experimental conditions

Table S1 shows the experimental conditions for the experiment presented in this study. Figure S1 shows the evolution of the α -pinene signal during the photooxidation.

Table S1: Experimental conditions for the experiment presented in this study.

VOC-to-NOx (ppbC/ppb)	NO (ppb)	NO2 (ppb)	OH exposure (#/cm ³ s)
7.4	22.5	58.3	2.34*10 ¹¹



15

Figure S1 Evolution of the α -pinene signal during the photooxidation.

S2 PMF details

S2.1 Minimum error for [PTRPTR-MS](#) data

Minimum error was determined by fitting a line for the last 1h of the experiment (see Fig. S2). During that time, only minor changes took place and the variation in the ion concentration was small. Therefore, this variation can be assumed to mostly consist of the noise in the data. The difference between the line fit and original signal (residuals) were calculated for each ion, and the standard deviation values were calculated from the residuals (see Fig. S3). Minimum error was selected as the median of those standard deviations. The minimum error was used to replace all values in the error matrix that were smaller than the minimum error. Figure S4 shows examples of the static error (a) and signal following error (b) described in the manuscript for the [PTRPTR-MS](#) data.

10

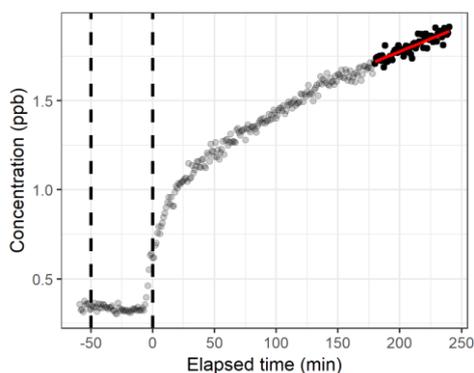
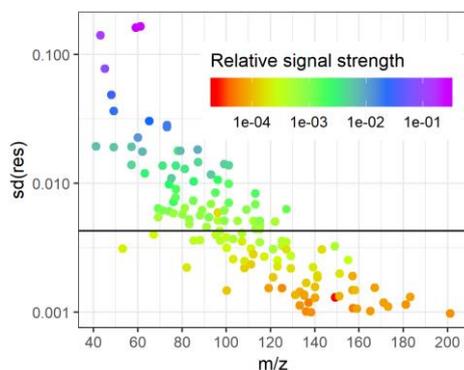


Figure S2 Example of the line fit (m/z 73.07, $C_4H_8O + H^+$). Dark points indicate the 1-hour period for which the linear fit (red) was applied.



15 **Figure S3** Standard deviations of the calculated residuals. Black horizontal line indicates the median value (0.0043) that is used as minimum error.

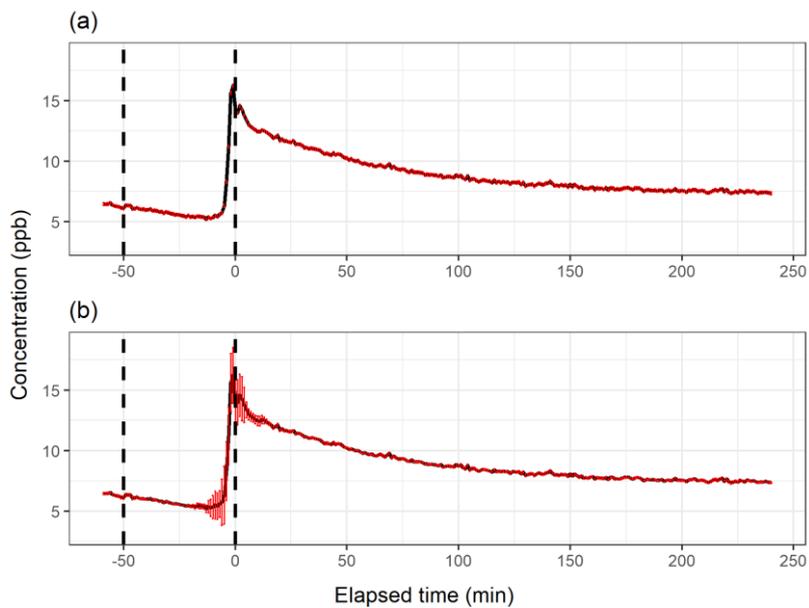


Figure S4 PMF error schemes for $C_2H_4O+H^+$ signal (m/z 45.03) from [PTR-PTR-MS](#) data. Static error in (a) and signal following error in (b).

S2.2 Minimum error for AMS data

- 5 Due to the small number of data points and slower reactions in the particle phase, the determination of the minimum error from the later parts of the experiment was not possible. Here, the minimum error was calculated as for the [PTR-PTR-MS](#) data, but the line fit was applied to the data before the photooxidation started (approximately 100 minutes). Before that time, no large changes in the particle mass concentration were observed, and the data mostly consisted on noise. Example of the line fit is shown in Fig. S5 and the standard deviations are shown in Fig. S6. Figure S7 shows an example of the static error
- 10 (a) and [Poisson type error standard AMS error](#) for AMS data.

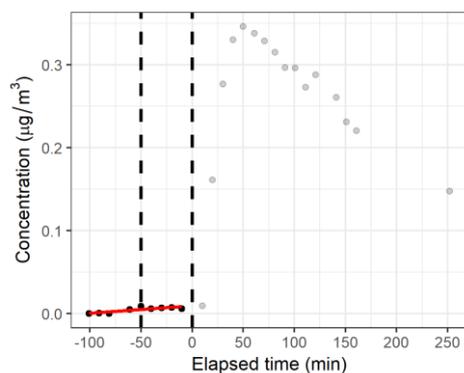


Figure S5 Example of the line fit (m/z 43.0548). Dark points indicate the first 100 minutes before the start of the photooxidation for which the linear fit (red) was applied.

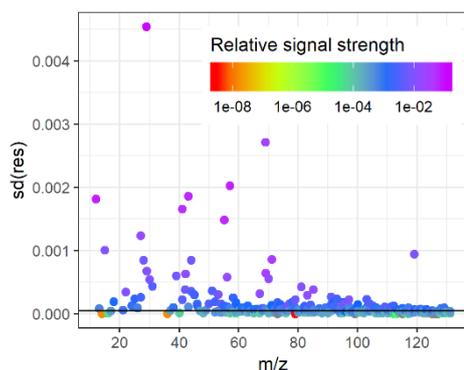
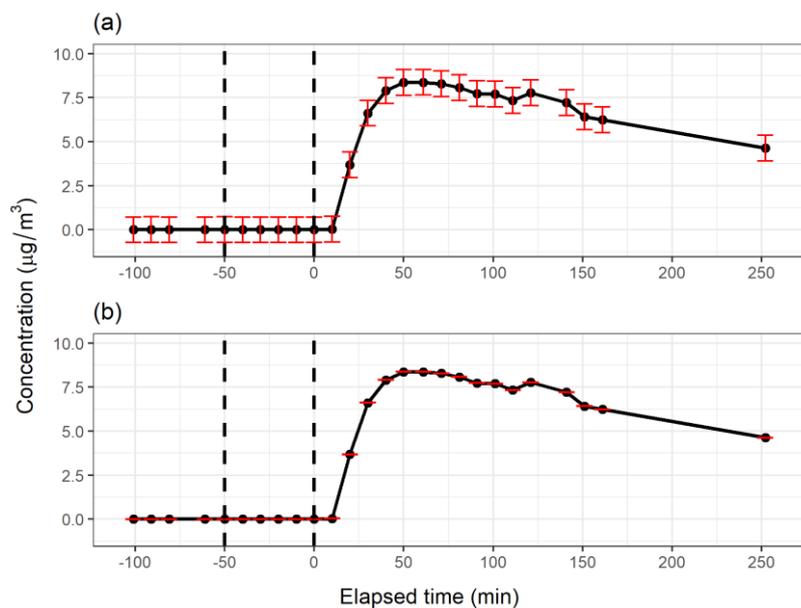


Figure S6 Standard deviations of the calculated residuals. Black horizontal line indicates the median value ($4.584 \cdot 10^{-5}$) that is used as a minimum error.



5

Figure S7 PMF error schemes for m/z 43.9898 from AMS data. Static error in (a) and Poisson-type error/standard AMS error in (b).

S2.3 Signal-to-noise ratios (SNR)

- 10 Signal-to-noise ratios were calculated in the Igor PMF toolkit (Ulbrich et al., 2009) for the data sets with different error schemes. SNR classifies ions either as “weak” ($0.2 < \text{SNR} < 2$) or “bad” ($\text{SNR} < 0.2$). For PTR-MS data with the static error, 12 ions out of 133 (9.0 %) were classified as weak and no bad variables ($\text{SNR} < 0.2$) were present. For the signal dependent error, 9 ions (6.8 %) were classified as weak. For AMS data with the static error, 8 ions out of 306 (2.6 %) were

classified as weak, and 1 ion (0.3 %) as bad. For the [Poisson-type error standard AMS error](#), 12 ions (3.9 %) were classified as weak, and 3 (0.98 %) as bad. However, no removal or downweighting of the variables was applied [in the results presented in the main manuscript](#), as for the weak ions the error values were already in the range of the ion signal itself, and for the bad variables the error was usually way above the signal throughout the ion time series. In addition, the number of weak or bad ions for both AMS and [PTRPTR-MS](#) data were rather small. [To justify our decision, we did run the PMF analysis with the downweighting. This, as expected, did not change our results or interpretation, and thus those results are omitted.](#)

S3 ~~Multivariate normality of the used data~~ [Various data statistics](#)

[S3.1 Multivariate normality of the used data](#)

Multivariate normality (MVN) of the [PTRPTR-MS](#) data was investigated with different tests presented detail in (Korkmaz et al., 2014). As a graphical approach, Fig. S8 shows the Chi-Square Quantile as a function of Mahalobnis distance (Q-Q plot). The points mainly follow the 1-1 reference line, but possible outliers are seen in the upper right corner. The other tests (Mardia's MVN test, Henze-Zirkler's test and Royston's MVN test) did not indicate multivariate normality. However, outlier values might have significant effect here and distort the test results (Korkmaz et al., 2014). For AMS data, the multivariate normality could not be stated due to the singularity issues in the calculation caused by the small data size (less rows than columns). When inspecting the univariate normality (Shapiro Wilk's tests) of the ions, none was declared as normally distributed.

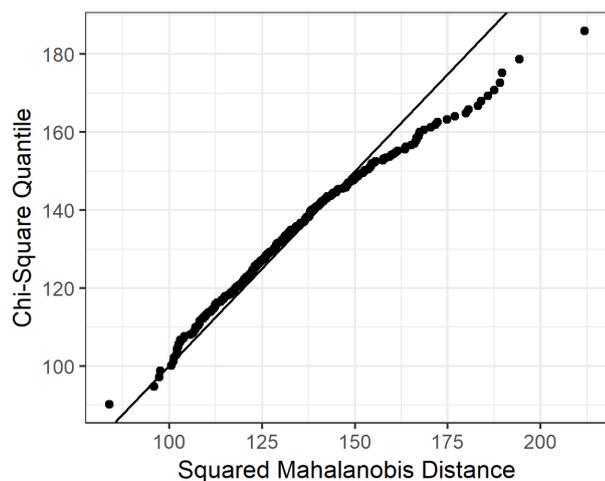


Figure S8 Chi-square quantile as a function of squared Mahalobnis distance for the measured [PTRPTR-MS](#) data.

[S3.2 Factor indexes for Exploratory Factor Analysis \(EFA\)](#)

[To investigate the number of factors for EFA, we calculated the standardized root mean residuals \(SRMR; Hu and Bentler, 1998\) and empirical Bayesian information criteria \(BIC; Schwarz, 1978\) -values. The SRMR can be considered as an](#)

average difference between the observed and reconstructed correlations, thus smaller values are preferred. It is calculated from the sum of squared residuals and adjusted for the degrees of freedom in the model. Given s_{ij} and σ_{ij} as an elements from the correlation matrix and reconstructed correlation matrix. . Hu and Bentler, 1998 give the following formulation:

$$SRMR = \sqrt{\frac{2 \sum_{i=1}^p \sum_{j=1}^i \left[\frac{s_{ij} - \sigma_{ij}}{s_{ii} s_{jj}} \right]^2}{p(p+1)}} \quad (S1)$$

- 5 The BIC, on the other hand, measures the balance of the fit between the increased likelihood (when adding parameters) of the model and a penalty term considering the number of parameters in the model (i.e. number of factors). In simplified format BIC can be presented as

$$BIC = \ln(n) k - 2 \ln(\hat{L}), \quad (S2)$$

- 10 when n is the number of data points, k is the number of parameters estimated by the model and \hat{L} is the maximized value of the likelihood function used inside the EFA algorithm.

S4 Additional SDRT results for PTRPTR-MS data

S4.1 EFA, PCA and PAM

- Figure S9 shows the 5-factor results from ml-EFA with Oblimin rotation and Fig. S10 shows the original loadings values from ml-EFA with Oblimin rotation as a scatter plots for the 4-factor solution presented in the manuscript. Figure S11 shows the explained variance for the number of components from SVD-PCA (applied to scaled data matrix) and the unrotated component time series and original loadings (scaled eigenvalues) are shown in Fig. S12 with 4 components. The unrotated results from SVD-PCA when applied to unscaled data matrix are shown in Fig. S13, and Fig. S14 shows the original loading values for the Oblimin rotated EVD-PCA with 4-components. Figure S15 and S16 shows results from PAM with 3 and 5 clusters, respectively. Table S2 presents the cluster sizes (number of compounds), maximum and average dissimilarities between the cluster compounds and the medoids, diameter of the cluster (maximum dissimilarity between two observations in a cluster) and the separation of the cluster (minimum dissimilarity between compounds in different clusters) for the 4-cluster solution presented in the manuscript. Table S2 shows the clustering statistics. Figure S17 shows the factor time series and contribution of ion to factor from dichotomized EFA (Oblimin rotated) with 4 factors.

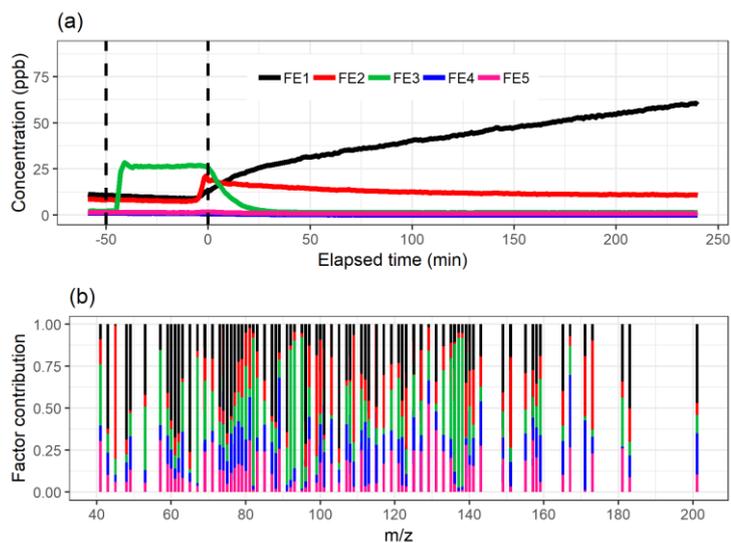
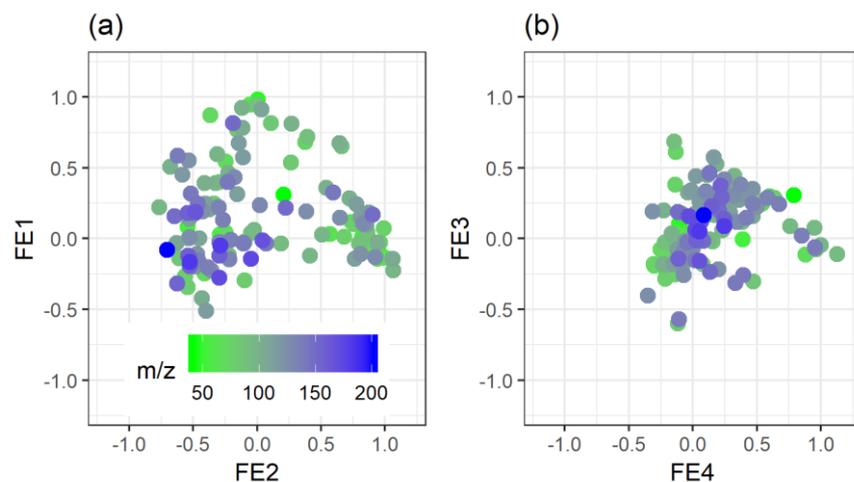


Figure S9 The factor time series (a) and total factor contribution (b) from ml-EFA with Oblimin rotation for the 5-factor solution. The colour code identifying factors is the same in both panels.



5

Figure S10 Unscaled factor loadings (4-factor solution) for the factors from ml-EFA with Oblimin rotation. The colour code is the same on both panels.

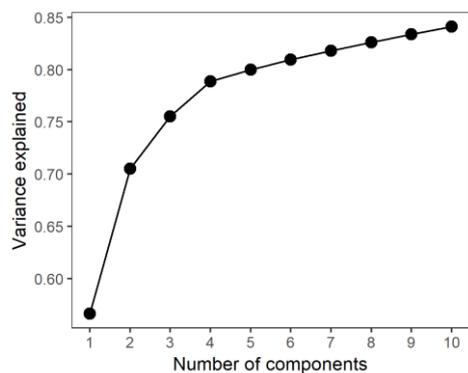
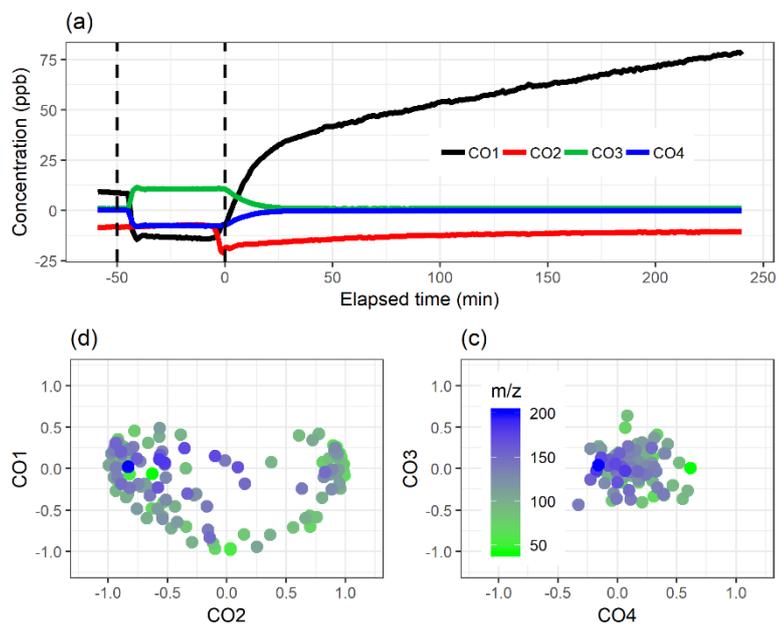


Figure S11 Explained variance as a function of number of components from SVD-PCA [applied to scaled PTR-MS data](#).



5

Figure S12 Unrotated component time series (a) and original loadings (b-c, scaled eigenvalues) from SVD-PCA [\(applied to scaled PTR-MS data\)](#) with 4 components. The colour code in (b) and (c) is the same.

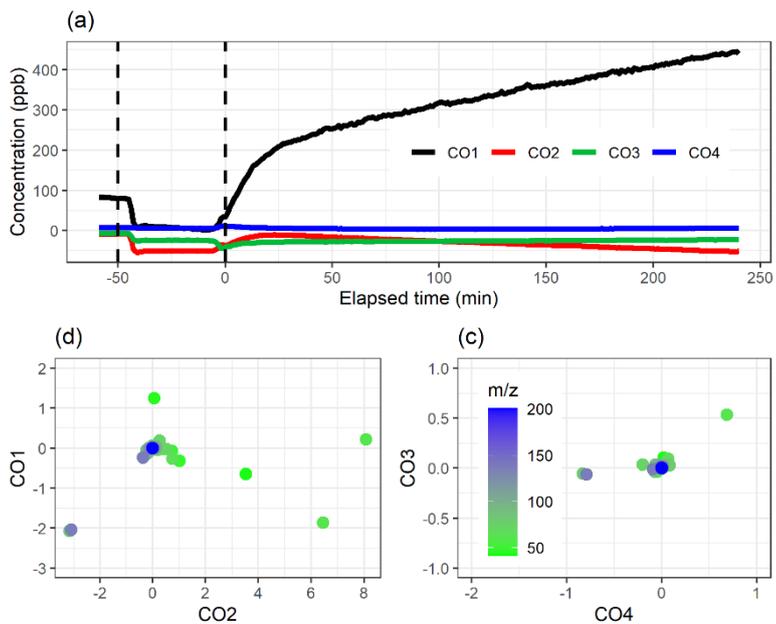


Figure S13 Unrotated component time series (a) and original loadings (b-c, scaled singular values) from SVD-PCA (applied to unscaled PTR-MS data) with 4 components. The colour code in (b) and (c) is the same.

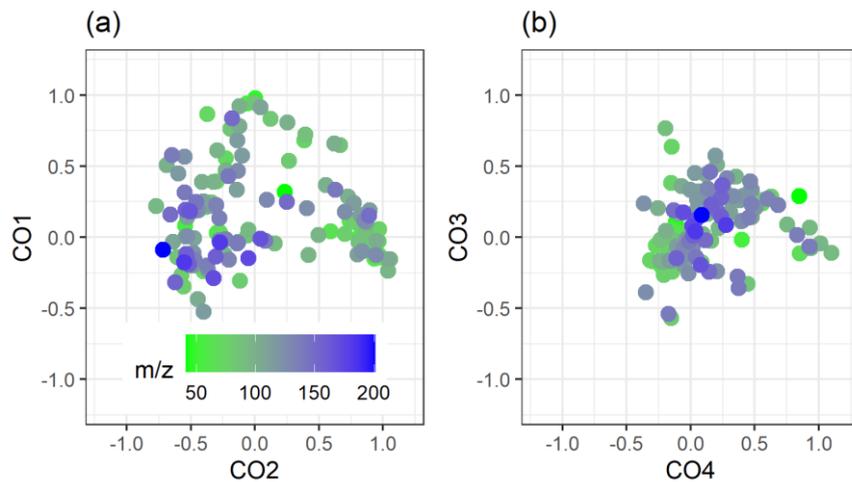


Figure S14 Unscaled component loadings (4-component solution) for the components from EVD-PCA with Oblimin rotation. The colour code is the same on both panels.

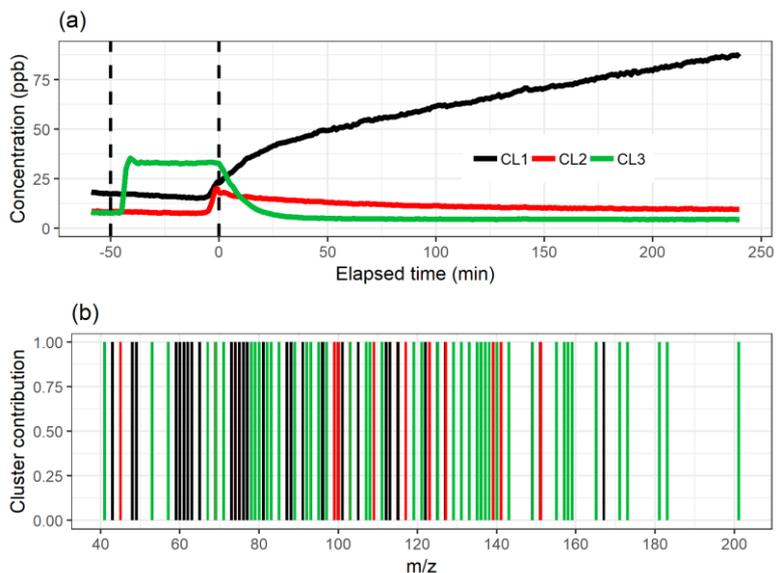
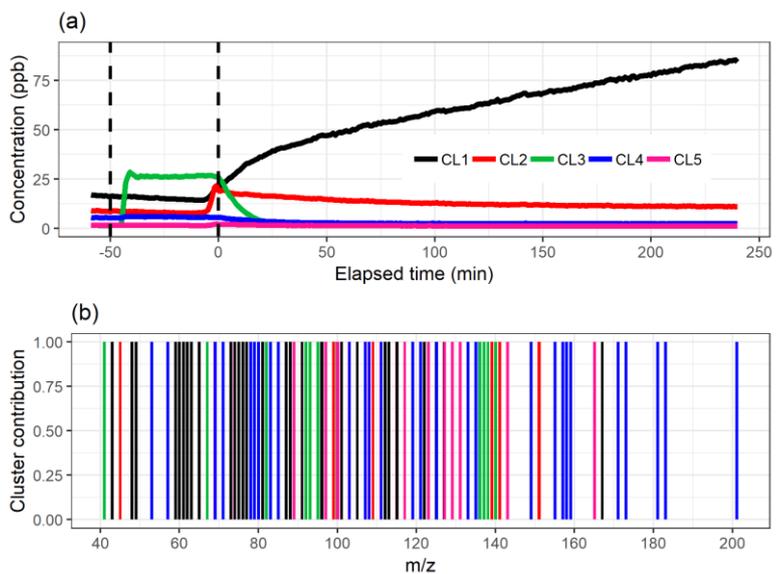


Figure S154 Cluster time series (a) and contribution of ion to cluster (b) from PAM with 3 clusters. The colour code identifying clusters is the same in both panels.

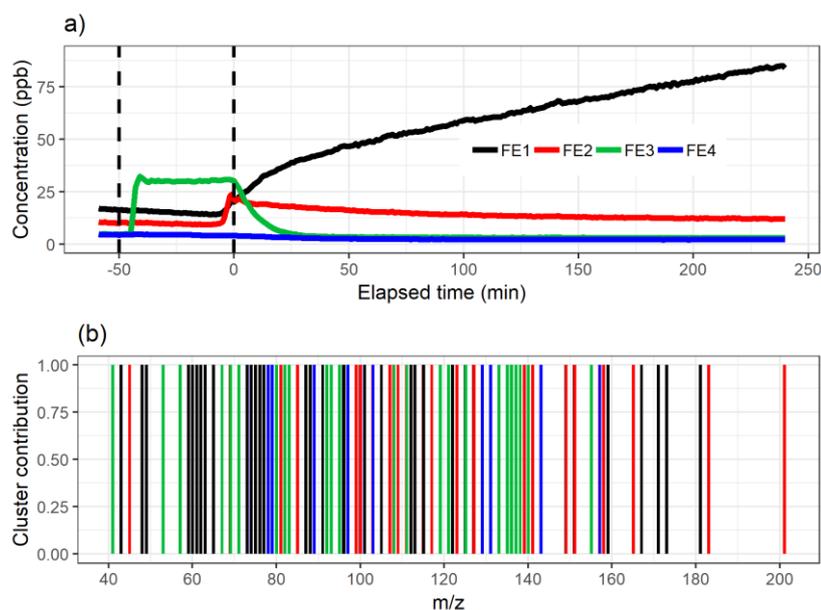


5

Figure S165 Cluster time series (a) and contribution of ion to cluster (b) from PAM for with 5 clusters. The colour code identifying clusters is the same in both panels.

Table S2 Clustering results for the measurement data with 4 clusters.

Cluster	Size	Maximum dissimilarity	Average Dissimilarity	Diameter	Separation	Medoid (m/z)
1	42	23.290	7.107	24.548	5.124	49.06 (C ₂ H ₈ O+H ⁺)
2	23	22.105	12.663	25.524	5.124	45.03 (C ₂ H ₄ O+H ⁺ , acetaldehyde)
3	14	22.288	5.587	24.188	5.307	137.13 (C ₁₀ H ₁₆ +H ⁺ , α -pinene)
4	54	23.121	11.284	24.890	5.307	107.09 (C ₈ H ₁₀ +H ⁺ , dimethylbenzene)



5

Figure S176 Factor time series (a) and contribution of ion to factor (b) from dichotomized EFA (Oblimin rotated) with 4 factors. The colour code identifying clusters is the same in both panels.

S4.2 NMF and PMF

Figure S186 shows the NMF results with only 4 factors. Figure S197 shows a boxplot of the distribution of the residuals for NMF with 4- and 5-factor solutions. Figure S2018 shows the PMF results with factorization rank 4 for the different error schemes.

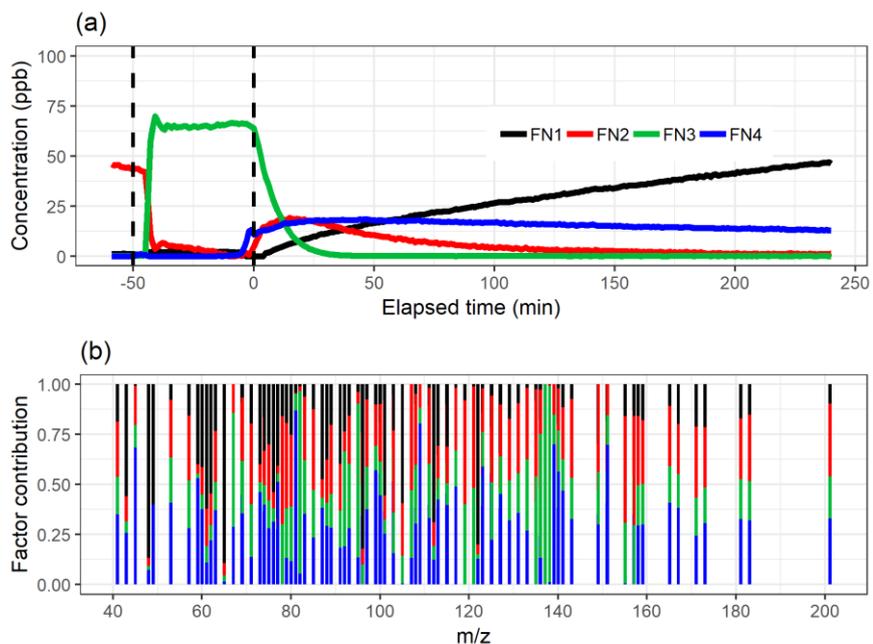
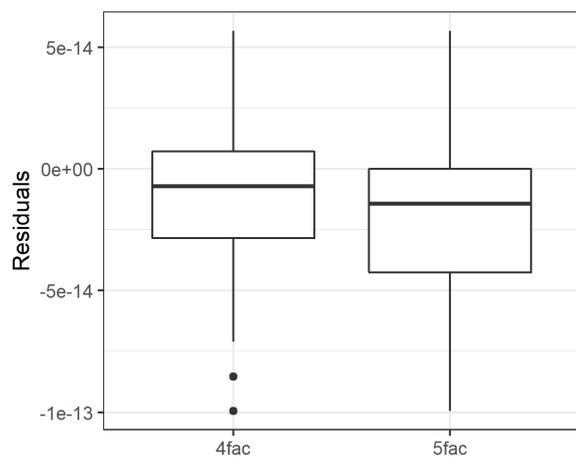


Figure S186 Factor time series (a) and contribution of ion to factor (b) from NMF with 4 factors. The colour code identifying clusters is the same in both panels.



5

Figure S197 Boxplot of the residuals (original total signal – reconstructed total signal) with 4 and 5 factors from NMF.

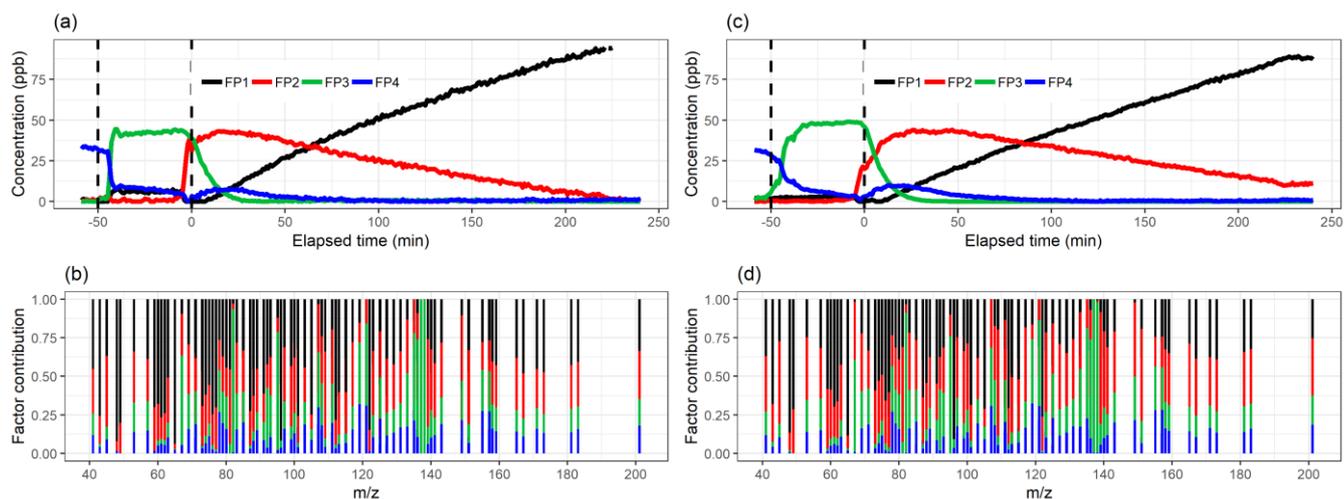


Figure S2018 Factor time series and contribution from PMF ($f_{\text{peak}} = 0$) with static error (a-b) and signal following error (c-d) for factorization rank 4. The colour code identifying the factors is the same in the top and bottom panels.

S5 Contrast angle [and factor spectra comparison](#)

- 5 The contrast angle describes how close two vector are in n -dimensional space. Here, n is the number of variables (ions) in the data. The contrast angle θ between two vectors can be calculated from

$$\cos(\theta) = \frac{(\vec{u} \cdot \vec{v})}{(|\vec{u}| \cdot |\vec{v}|)}, \quad (\text{S34})$$

where u and v are the two vectors to be compared. The distribution of ions (see e.g. Fig. 3b) into the different factors between different SDRTs can be then compared by calculating the contrast angle between the same factor acquired with different methods (i.e. between factor1 from EFA and NMF, Figs. 3b and 6b). The larger the contrast angle is ($\theta = [0, 90)$), the farther apart the factors are from each other (i.e. more different) in the n -dimensional space. The contrast angles between the SDRTs, excluding PAM, for the [PTR-MS](#) data are shown in Table S3.

The contrast angles between EFA and PCA are very small, indicating the distribution of ions between the factors in these methods is indeed similar, as also noted in the manuscript. Differences between the different error schemes for PMF are also rather small, however, for factor2 the difference is slightly larger. The differences are obviously larger when the methods had different number of factors (4 factors were selected for PCA and EFA, and 5 factors for PMF and NMF).

[Figure S21 and S22 show the factor contributions in separate panels for each factor for the main results from gas phase data \(PTR-MS\). Despite the fundamental differences between these methods, the relative factor contributions clearly show similarity. Even PAM, that is relatively different when compared to other SDRTs as it only assigns one m/z into one factor, the same ions are enhanced when compared to other SDRTs.](#)

[Table S4 shows the contrast angles for AMS data between NMF and PMF with static error scheme. PMF with Poisson style error is omitted due to the inability to interpret the factors reasonably. For the AMS data the factors were rather different](#)

when inspecting the factor time series. However, the distribution of ions seemed to agree decently. From Table S4 we may note, that between factors 2-4 the contrast angle is indeed small, indicating good agreement between the SDRTs. However, for the factor1 (interpreted as primary OA, HOA) the contrast angle is larger, indicating a larger number of ions have different contribution to factor1 between NMF and PMF. Indeed, when we inspect the larger m/z in Figs. 12b and 13b, PMF has more contribution of ions with larger m/z compared to NMF.

Table S3 Contrast angles for the factors from SDRTs applied to the measured PTR-MS data. PMF1 refers to PMF with static error scheme, and PMF2 for the signal following error.

factor	1	2	3	4	5	SDRT-pair
θ (°)	3.0	13.0	6.7	1.5	2.8	PMF1/PMF2
	16.7	13.8	3.8	4.5	20.8	PMF1/NMF
	19.0	20.6	8.6	3.7	19.4	PMF2/NMF
	30.5	31.7	23.4	35.0	NA	EFA/NMF
	30.6	31.9	24.4	35.4	NA	PCA/NMF
	23.8	33.5	25.6	36.6	NA	PMF1/EFA
	23.6	35.3	27.8	36.1	NA	PMF2/EFA
	24.1	33.7	26.7	37.1	NA	PMF1/PCA
	23.9	35.5	29.2	36.5	NA	PMF2/PCA
	1.9	1.9	3.5	3.1	NA	PCA/EFA

Table S4 Contrast angles for the factors from PMF and NMF applied to the measured AMS data. PMF1 refers to PMF with static error scheme.

factor	1	2	3	4	SDRT-pair
θ (°)	51.0	4.5	2.6	6.2	PMF1/NMF

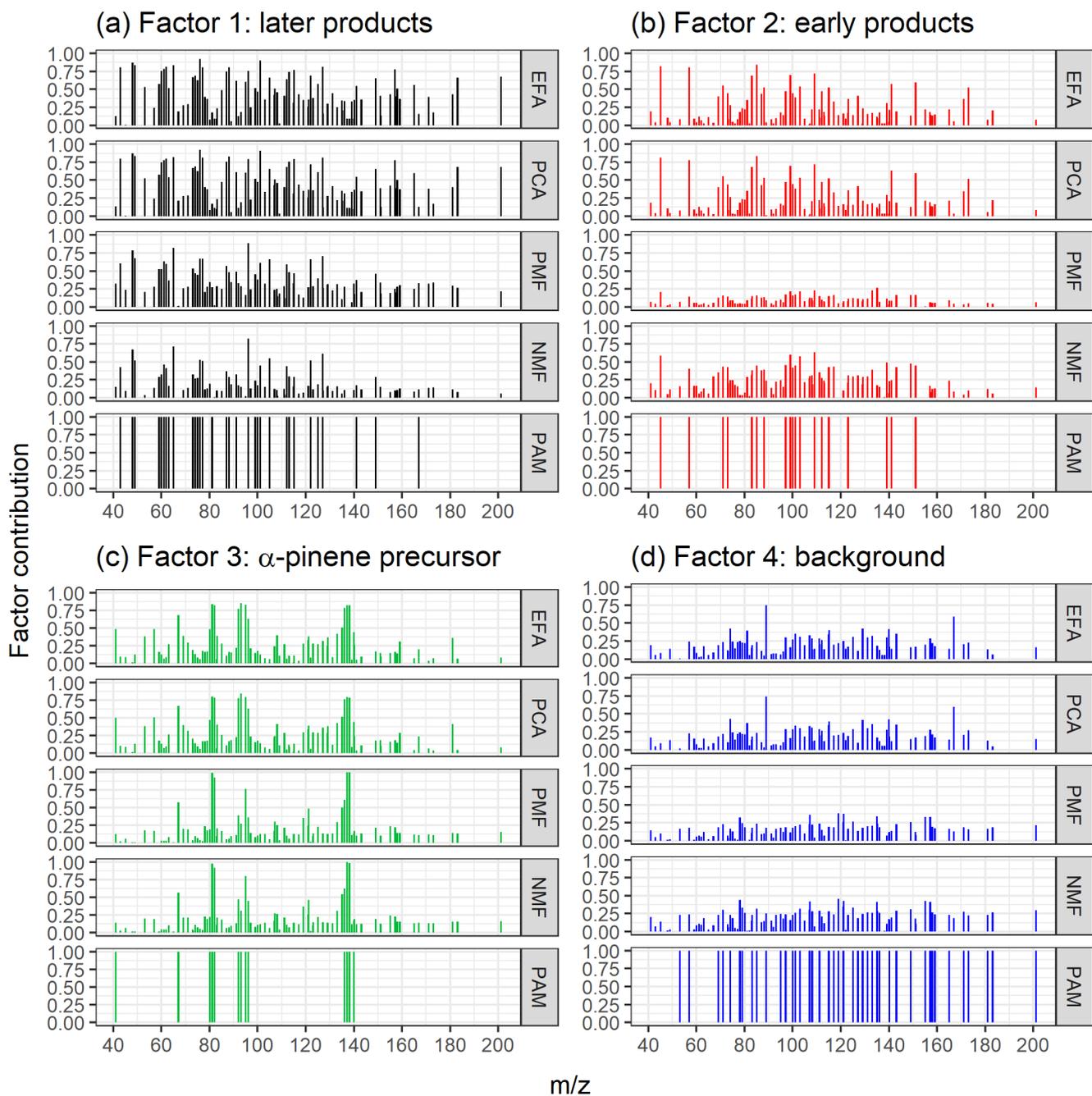
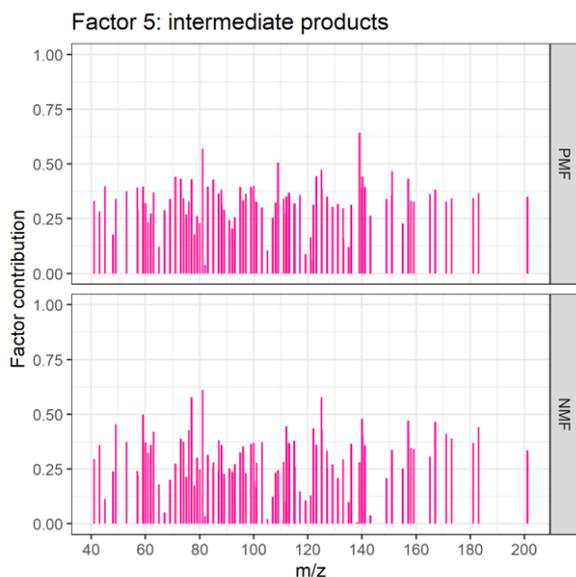


Figure S21 Factor contribution for factors 1-4 from ml-EFA, EVD-PCA, PAM, NMF and PMF with static error for the gas phase data measured with PTR-MS.



[Figure S22](#) Factor contribution for factor 5 from NMF and PMF with static error for the gas phase data measured with PTR-MS.

S6 Additional SDRT results for AMS data

S6.1 EFA, PCA and PAM

- 5 Figure [S19S23](#) shows the test results for the different number of factors, components and clusters for EFA, PCA and PAM, respectively. Figures [S240](#) and [S254](#) show pa-EFA results for the AMS data with 2 and 3 factors. More factors were also tested, but only the 2 factors can be separated. PCA (EVD and SVD) had similar behaviour, [and thus not shown here](#). In addition, the loading value distribution between the first two factors in Fig. [S240b](#) barely changes when adding a new factor (Fig. [S254b](#)), indicating the algorithm is not able to separate any additional factors from the first 2. Figures [S262](#) and [S273](#) show the results from PAM with 2-5 clusters.
- 10

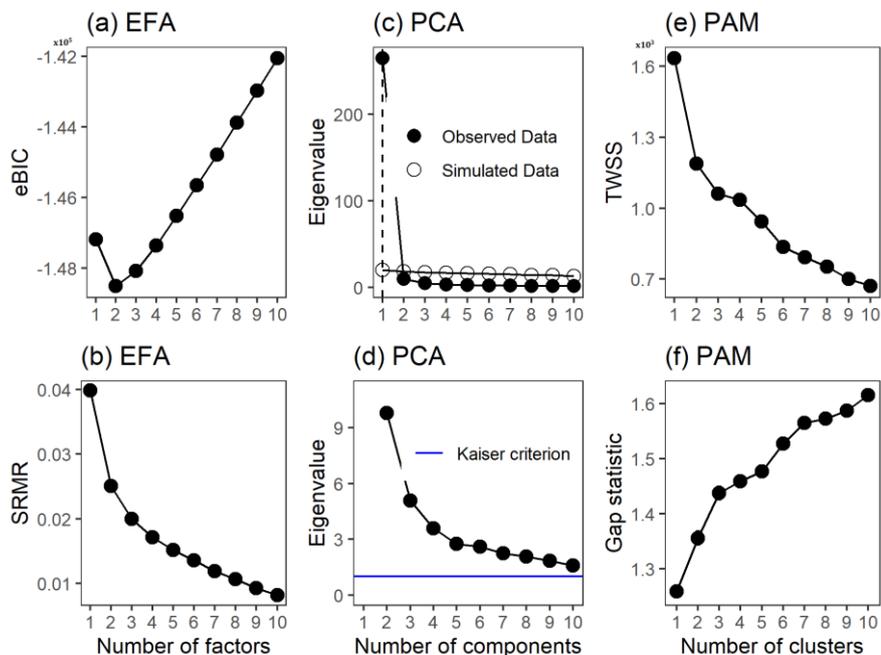
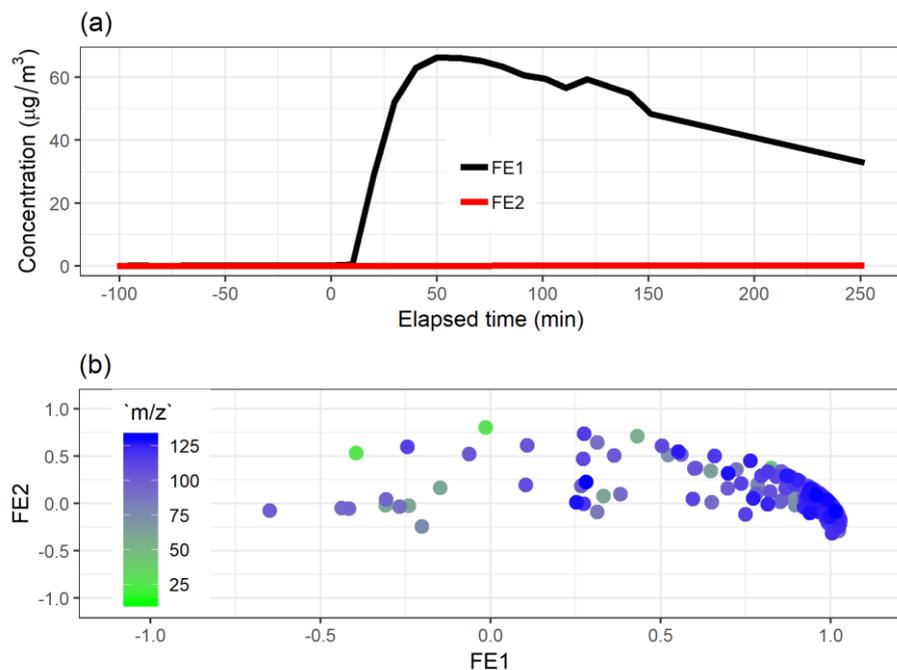


Figure S19S23 Factor number indexes for particle phase data (AMS). Empirical BIC (a) and SRMR (b) as a function of the number of factors for pa-EFA, Parallel analysis (c) and Kaiser criterion (d) for EVD-PCA and TWSS (e) and Gap statistic (f) for PAM. Data measured with AMS.



5

Figure S240 Oblimin rotated factor time series (a) and original loadings (b) from pa-EFA with 2 factors for the particle phase data measured with AMS.

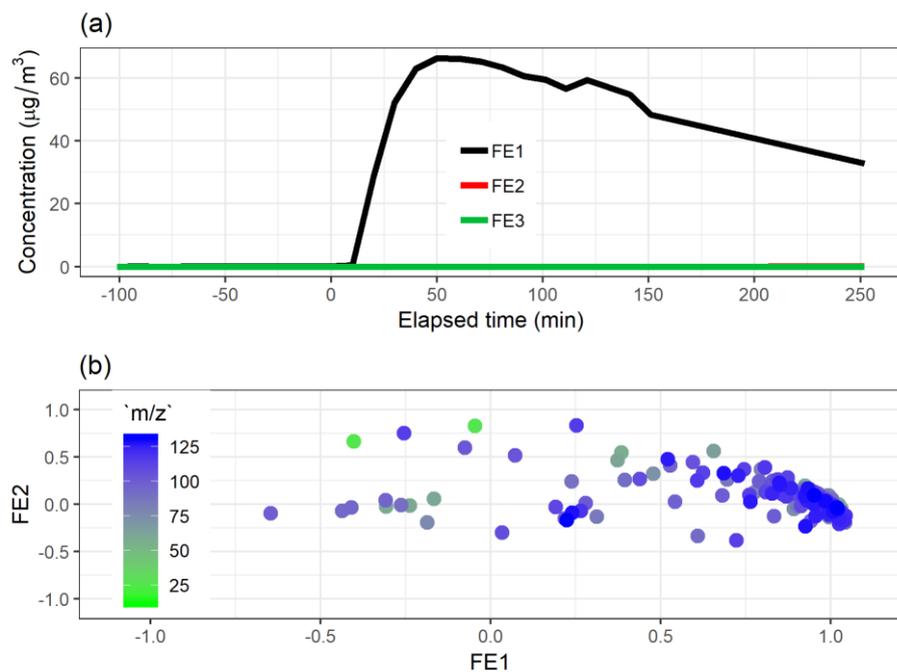
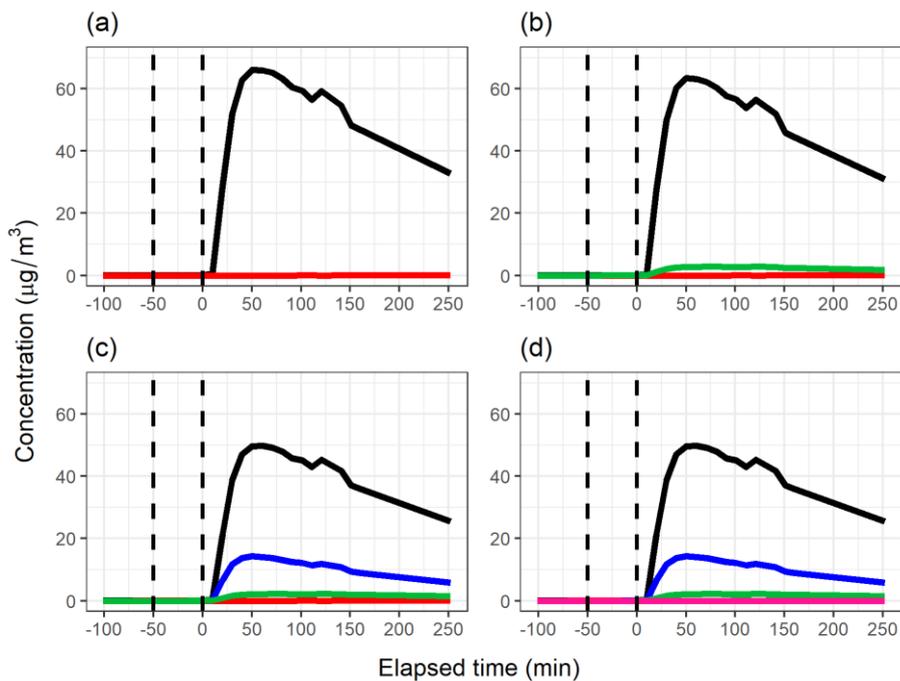


Figure S251 Oblimin rotated factor time series (a) and original loadings (b) from pa-EFA with 3 factors for the [particle phase](#) data measured with AMS.



5 Figure S262 Cluster time series from PAM with (a) 2, (b) 3, (c) 4 and (d) 5 clusters for the measured [particle phase \(AMS\)](#) data.

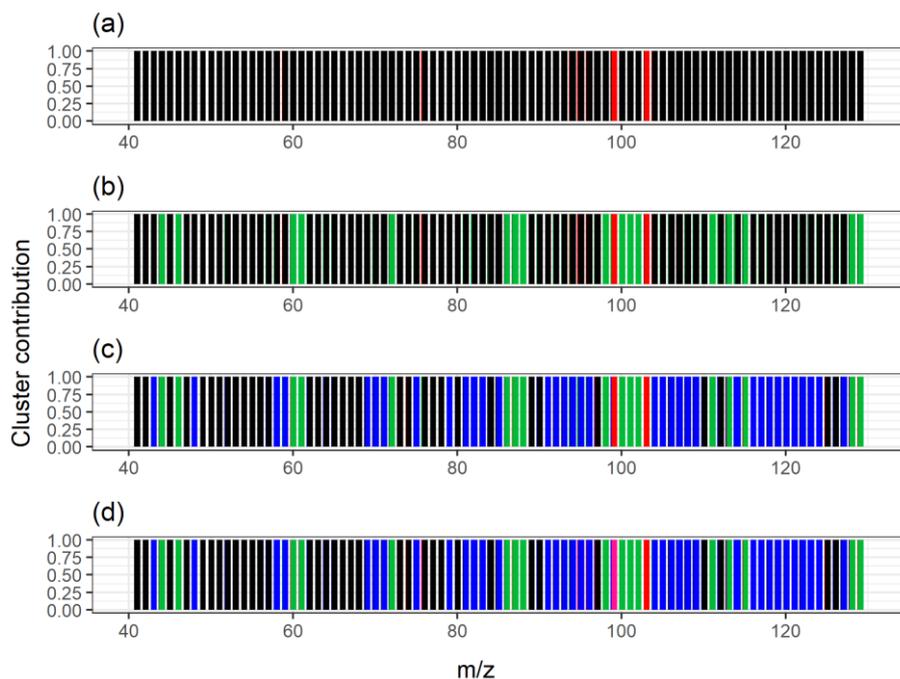


Figure S273 Cluster contribution from PAM with (a) 2, (b) 3, (c) 4 and (d) 5 clusters for the measured [particle phase \(AMS\)](#) data. The colours in the subplots correspond to the time series in [Fig. S262](#). Ions with $m/z > 40$ are omitted from the plot as no changes were observed (ions were in black factor in all cases).

5 S6.2 NMF and PMF

To further justify the selected number of factors in NMF for AMS data, we compared the results from the 3-, 4- and 5-factor solutions. In the 3-factor solution ([Fig. S283](#)), FN1 appeared before $t = 0$ min indicating it includes mostly primary OA. However, this factor decreased to 0 at the start of photooxidation, but then increased again around $t = 80$ min. In the 4-factor case, the corresponding factor exhibits a small increase at $t = 0$ min and stays rather constant after that. In addition, with 3 factors the LVOOA factor (FN3 in the 4-factor case) is not observed at all, indicating 3 factors is not enough to separate LVOOA from the primary HOA. The decision between 4-factor ([Fig. 12 in the main text](#)) and 5-factor ([Fig. S294](#)) solutions is more difficult. FN3 is the same in both cases. FN1 which must be identified as containing “primary” OA from the car emissions as it captures the signal at $t < 0$ min, shows opposite behaviour at the onset of photooxidation. In the 4-factor solution it increases, but in the 5-factor solution it drops to 0 in less than 10 min. This seems to be unlikely behaviour, as no such sudden loss process is expected in the particle phase. The main reason for decreasing signals in the AMS is the overall particle wall loss, which will affect all compounds equally as long as there is no strong particle size dependence of the particle composition. Other reasons for individual compounds decreasing are evaporation of semi volatile compounds if their concentration in the gas phase changes or particle phase chemistry. Although these processes may be started by the onset of photochemistry in the gas phase, neither of them is expected to be that fast. The other main impact of the additional factor is a redistribution of the signal in FN2 and FN4 (4-factor solution) to factors FN2, FN4 and FN5 (5-factor solution).

The stable concentration value of FN5 after the initial fast increase can only be explained by a constant source for these compounds large enough to compensate the overall particle loss to the chamber walls. Again, this seems rather unlikely for experiment conditions. Overall, the 4-factor solution has the most interpretable results and the statistical tests suggest it as a good solution.

- 5 Figures [from S3026 to S32](#) and [S27](#) show the PMF results with 3 [to](#)and 5 factors, [respectively](#).

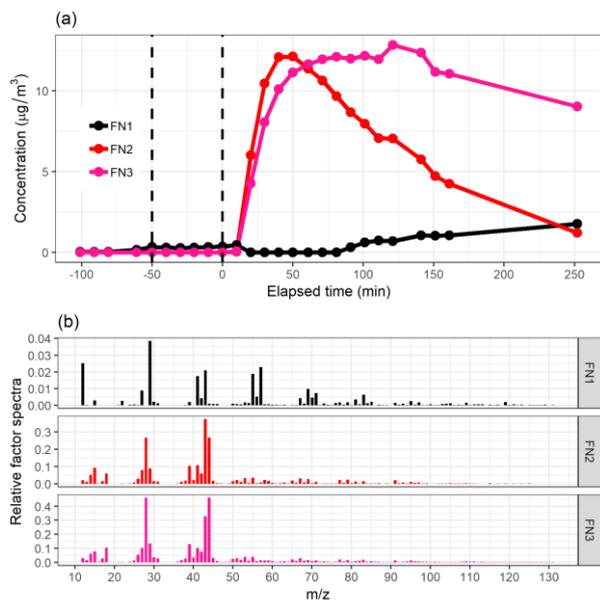


Figure S284 Factor time series (a) and relative factor spectra (b) from NMF with 3 factors for the measured [particle phase \(AMS\)](#) data. The colour code identifying the factors is the same in both panels.

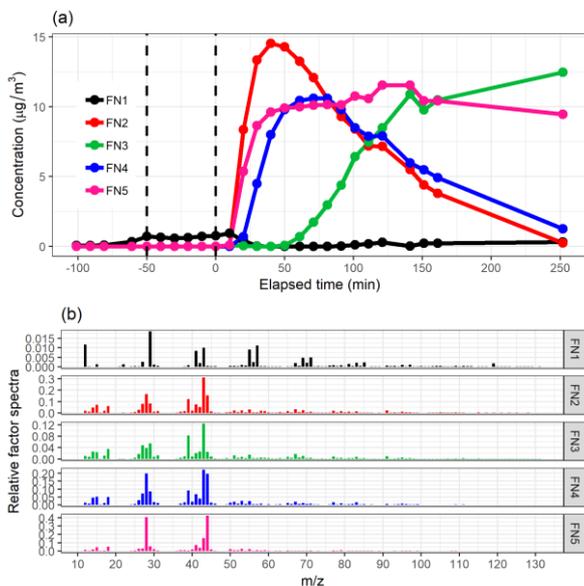
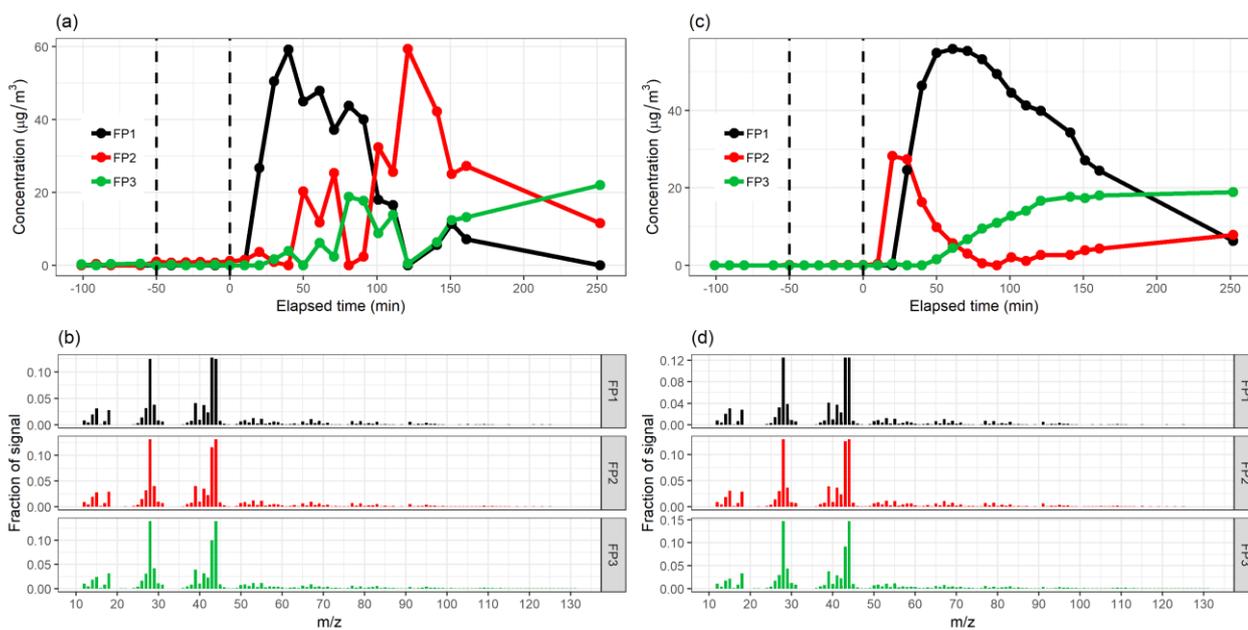


Figure S295 Factor time series (a) and relative factor spectra (b) from NMF with 5 factors for the measured [particle phase \(AMS\)](#) data. The colour code identifying the factors is the same in both panels.



5 Figure S3026 Factor time series and contribution from PMF with static error (a-b) and [Poisson-style error](#) [Standard AMS error](#) (c-d) for factorization rank 3 for the measured [particle phase \(AMS\)](#) data.

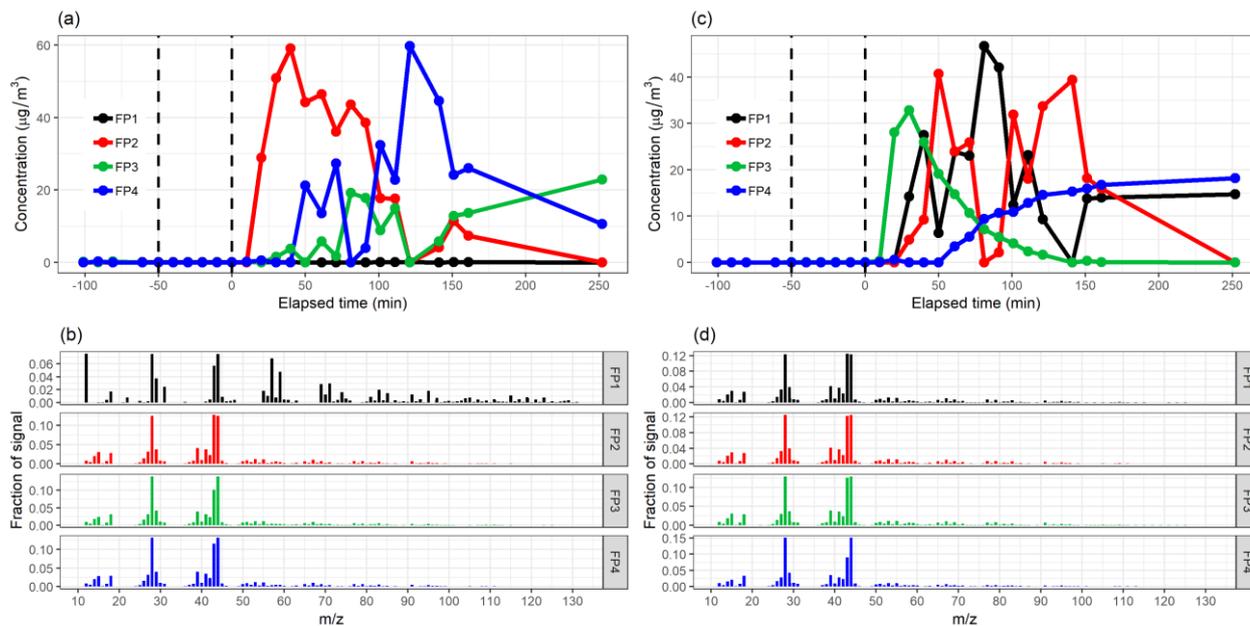
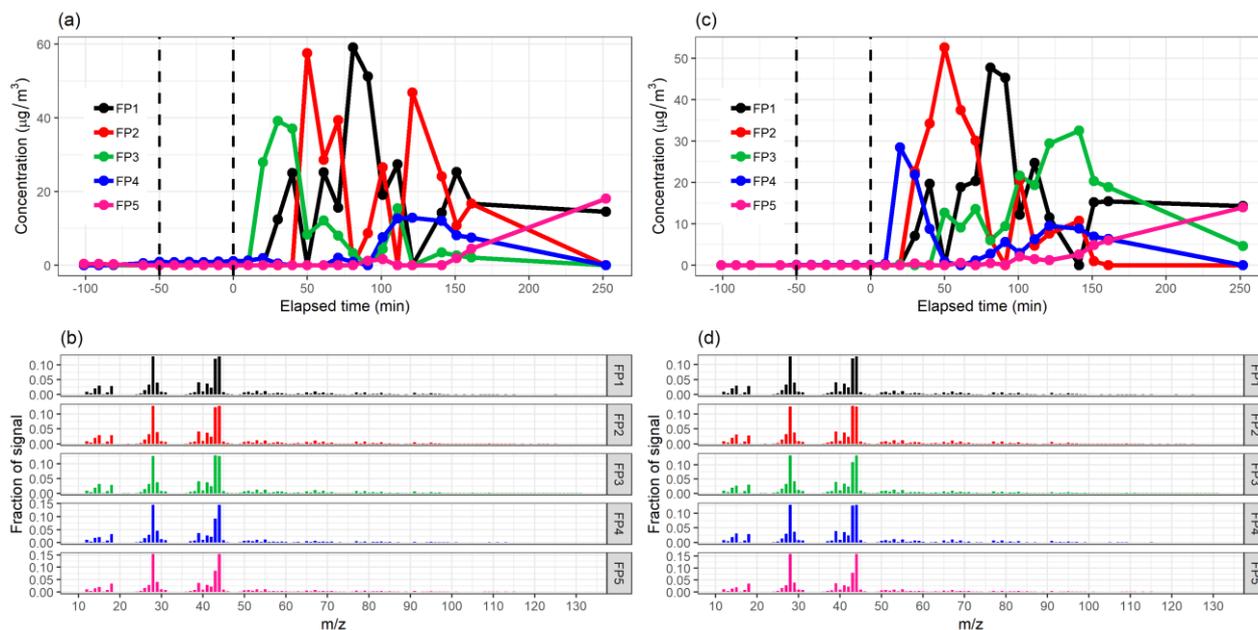


Figure S31 Factor time series and contribution from PMF with static error (a-b) and Standard AMS error (c-d) for factorization rank 4 for the measured particle phase (AMS) data.



5

Figure S327 Factor time series and contribution from PMF with static error (a-b) and Poisson-style error Standard AMS error (c-d) for factorization rank 5 for the measured particle phase (AMS) data.

S7 References

Hu, L. T., and Bentler, P. M.: Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification, *Psychol Methods*, 3, 424-453, Doi 10.1037/1082-989x.3.4.424, 1998.

5 Korkmaz, S., Goksuluk, D., and Zararsiz, G.: MVN: An R Package for Assessing Multivariate Normality, *R J*, 6, 151-162, 2014.

Schwarz, G.: Estimating the Dimension of a Model, *Annals of Statistics*, 6, 461-464, doi:10.1214/aos/1176344136, 1978.

10 Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, *Atmos Chem Phys*, 9, 2891-2918, 10.5194/acp-9-2891-2009, 2009.