Supplement of Atmos. Meas. Tech., 13, 2995–3022, 2020
https://doi.org/10.5194/amt-13-2995-2020-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

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

Sini Isokääntä et al.

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

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}

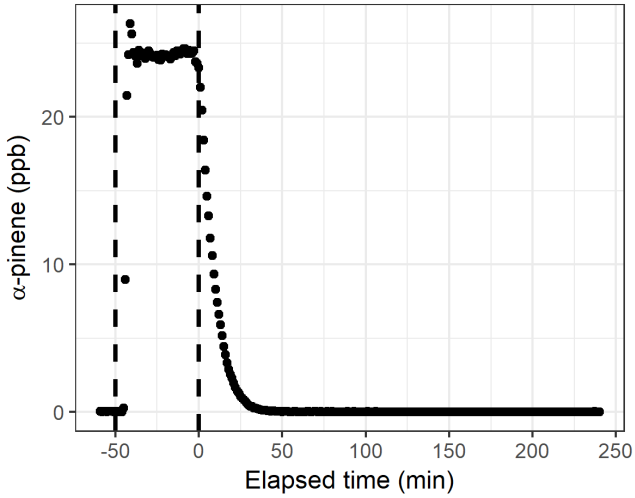

Figure S1 Evolution of the α-pinene signal during the photooxidation.

S2 PMF details

S2.1 Minimum error for PTR-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 PTR-MS data.

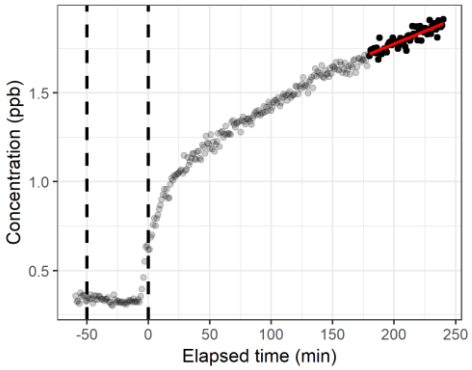

Figure S2 Example of the line fit (m/z 73.07, $C_4H_8O + H^+$). Dark points indicate the 1-hour period for which the linear fit (red) was applied.

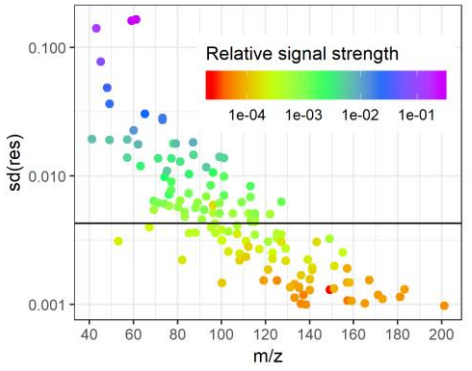

Figure S3 Standard deviations of the calculated residuals. Black horizontal line indicates the median value (0.0043) that is used as minimum error.

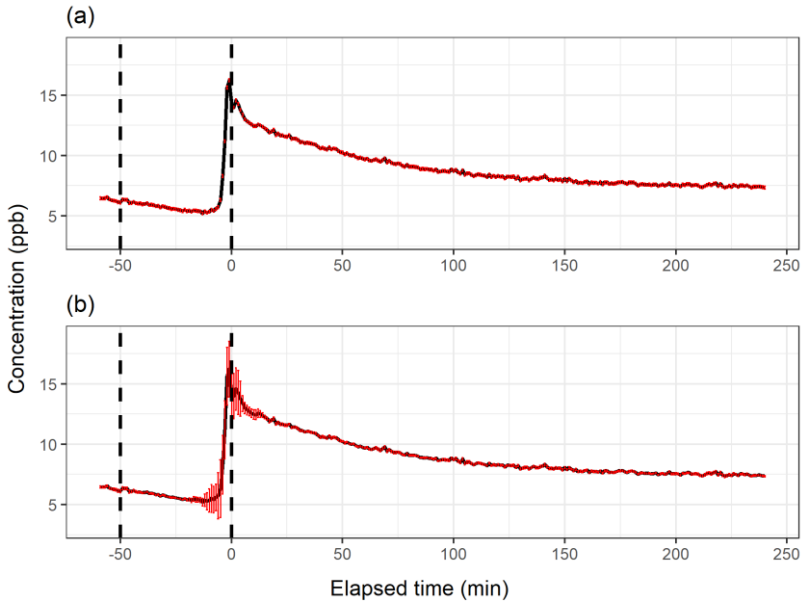

Figure S4 PMF error schemes for $C_2H_4O+H^+$ signal (m/z 45.03) from PTR-MS data. Static error in (a) and signal following error in (b).

S2.2 Minimum error for AMS data

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-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 (a) and Standard AMS error for AMS data.

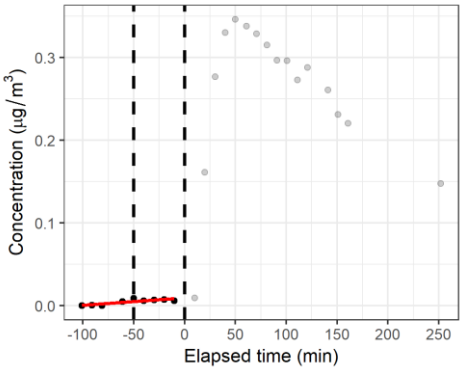

Figure S5 Example of the line fit (m/z 43.0548). Dark points indicate the first 100 minutes before the start of the photooxidation for which the linear fit (red) was applied.

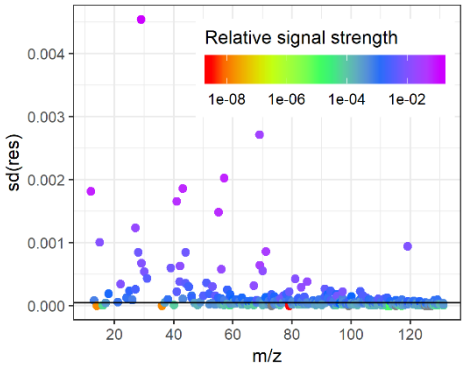

Figure S6 Standard deviations of the calculated residuals. Black horizontal line indicates the median value ($4.584 \cdot 10^{-5}$) that is used as a minimum error.

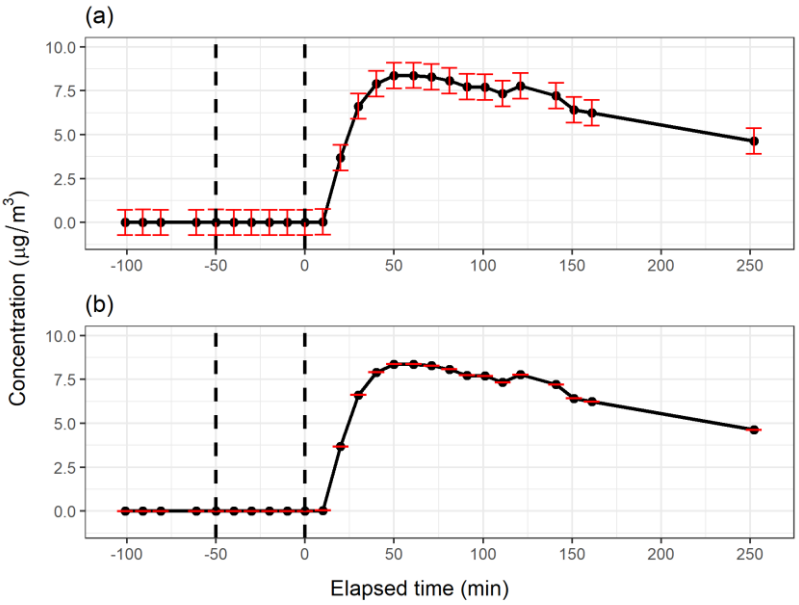

Figure S7 PMF error schemes for m/z 43.9898 from AMS data. Static error in (a) and Standard AMS error in (b).

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

Signal-to-noise rations 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 < SNR < 2) or "bad" (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 (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 Standard AMS error, 12 ions (3.9 %) were classified as weak, and 3 (0.98 %) as bad.

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.

S3 Various data statistics

S3.1 Multivariate normality of the used data

Multivariate normality (MVN) of the PTR-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.

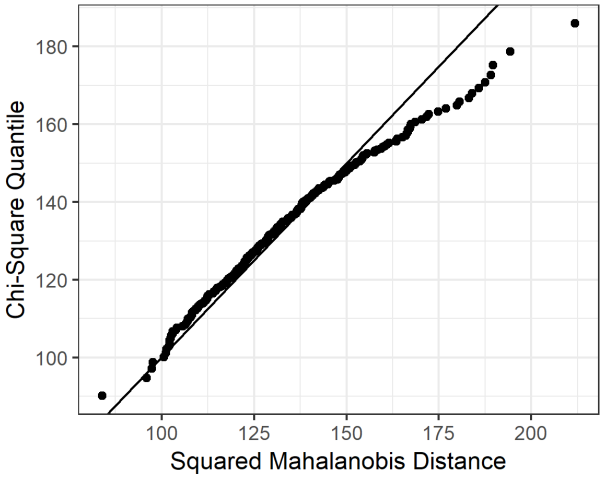

Figure S8 Chi-square quantile as a function of squared Mahalobnis distance for the measured PTR-MS data.

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;

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

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

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

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 PTR-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. Figure S17 shows the factor time series and contribution of ion to factor from dichotomized EFA (Oblimin rotated) with 4 factors.

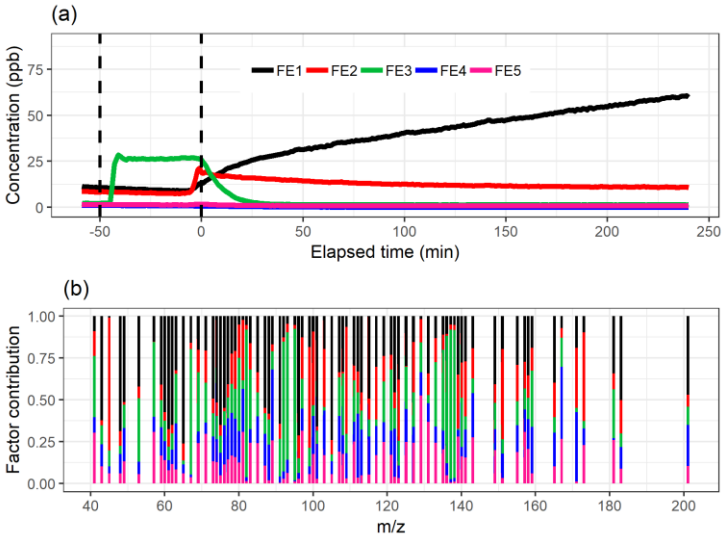

Figure S9 The factor time series (a) and total factor contribution (b) from ml-EFA with Oblimin rotation for the 5-factor solution. The colour code identifying factors is the same in both panels.

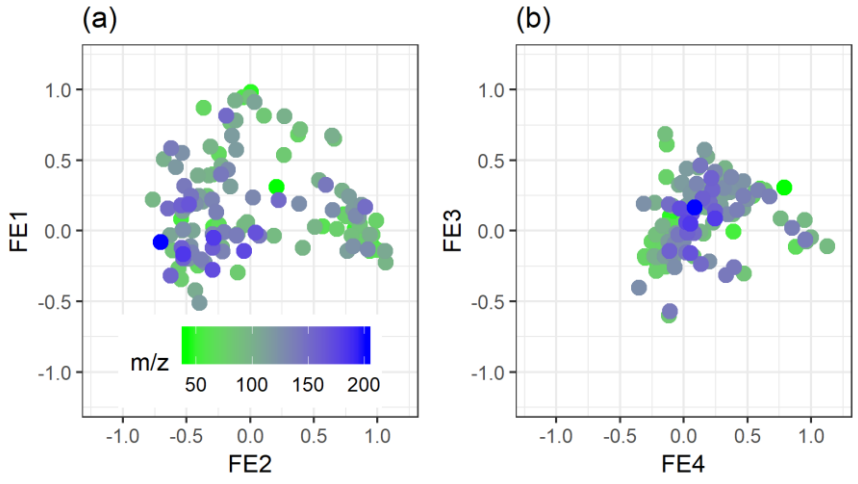

Figure S10 Unscaled factor loadings (4-factor solution) for the factors from ml-EFA with Oblimin rotation. The colour code is the same on both panels.

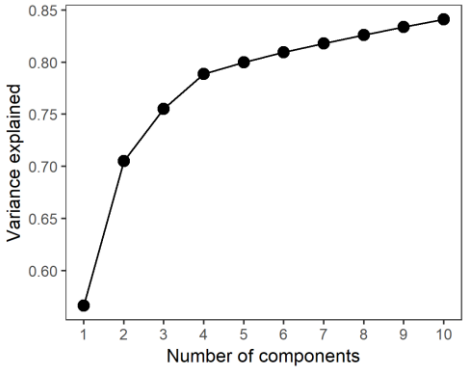

Figure S11 Explained variance as a function of number of components from SVD-PCA applied to scaled PTR-MS data.

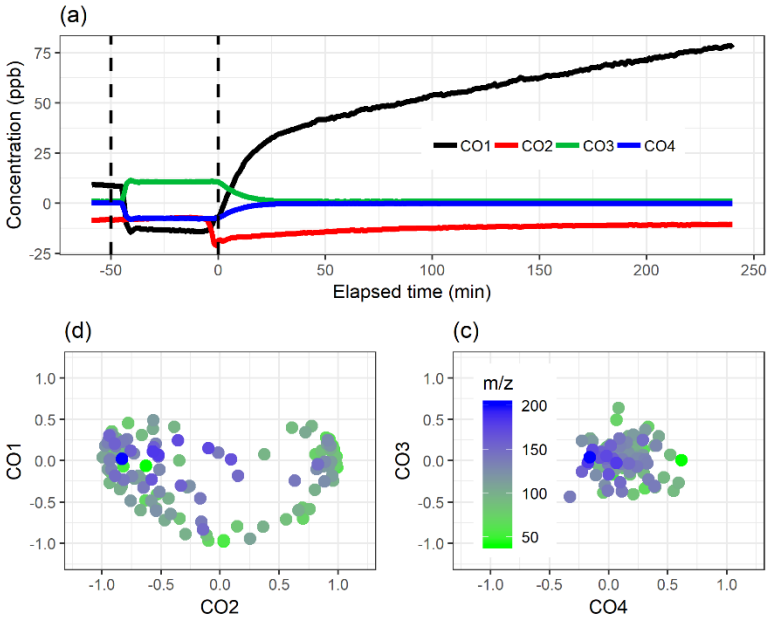

Figure S12 Unrotated component time series (a) and original loadings (b-c, scaled eigenvalues) from SVD-PCA (applied to scaled PTR-MS data) with 4 components. The colour code in (b) and (c) is the same.

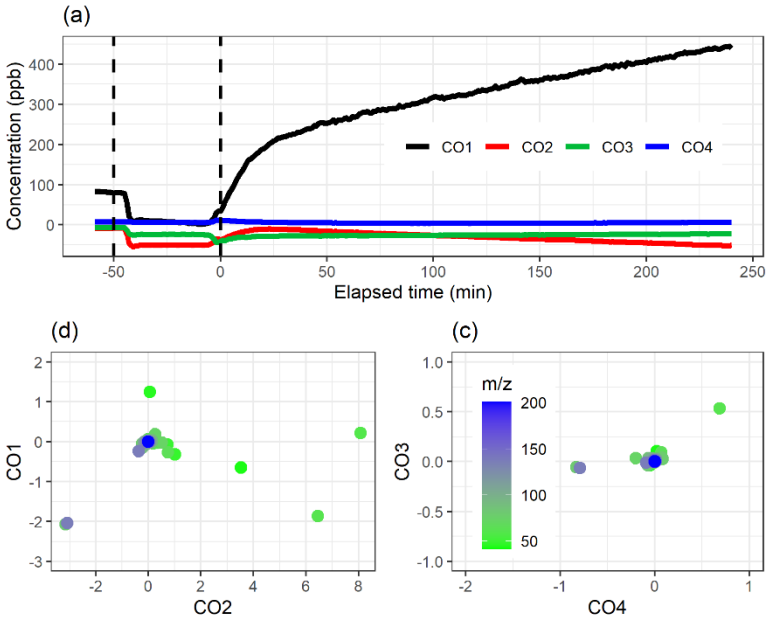

Figure S13 Unrotated component time series (a) and original loadings (b-c, scaled singular values) from SVD-PCA (applied to unscaled PTR-MS data) with 4 components. The colour code in (b) and (c) is the same.

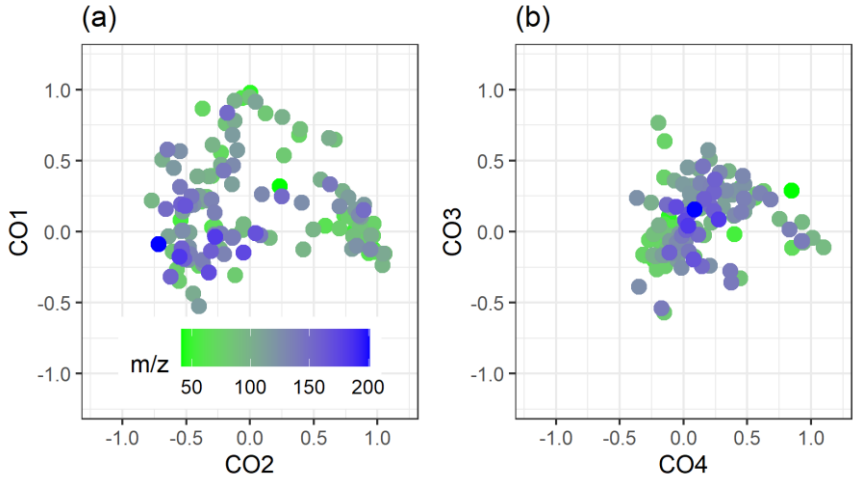

Figure S14 Unscaled component loadings (4-component solution) for the components from EVD-PCA with Oblimin rotation. The colour code is the same on both panels.

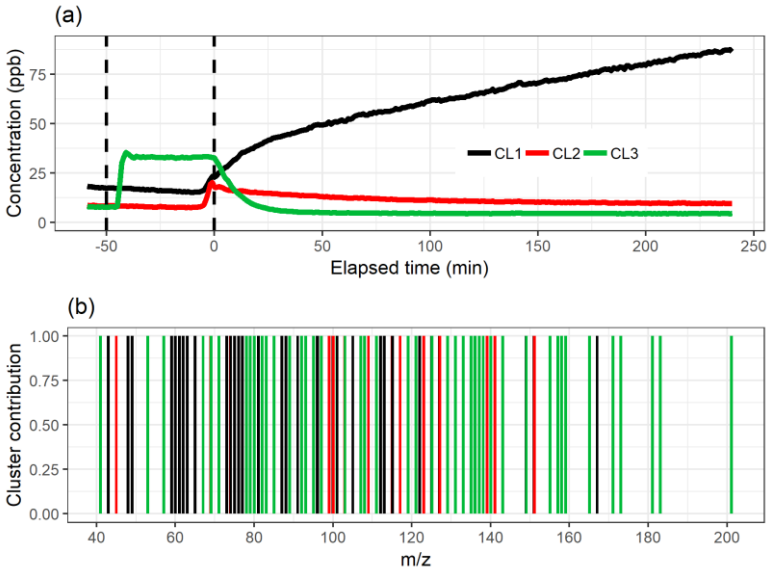

Figure S15 Cluster time series (a) and contribution of ion to cluster (b) from PAM with 3 clusters. The colour code identifying clusters is the same in both panels.

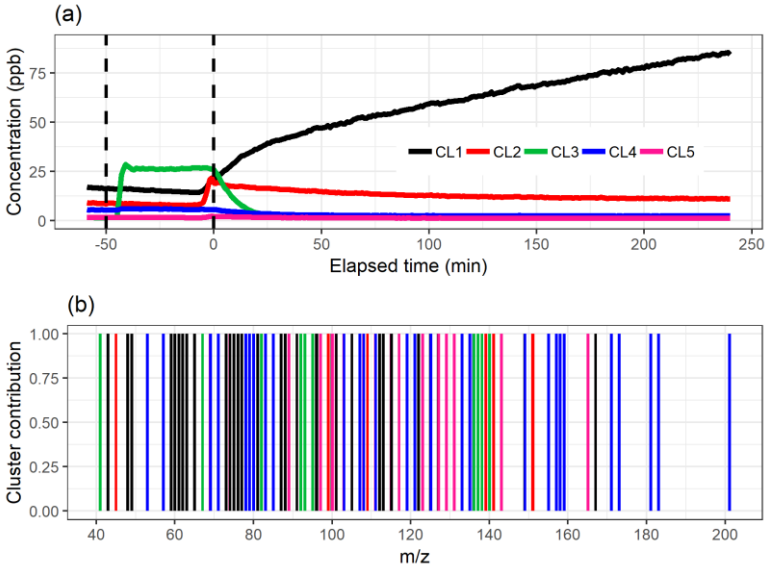

Figure S16 Cluster time series (a) and contribution of ion to cluster (b) from PAM for with 5 clusters. The colour code identifying clusters is the same in both panels.

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 ($C_2H_8O+H^+$)
2	23	22.105	12.663	25.524	5.124	45.03 ($C_2H_4O+H^+$, 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 ($C_8H_{10}+H^+$, dimethylbenzene)

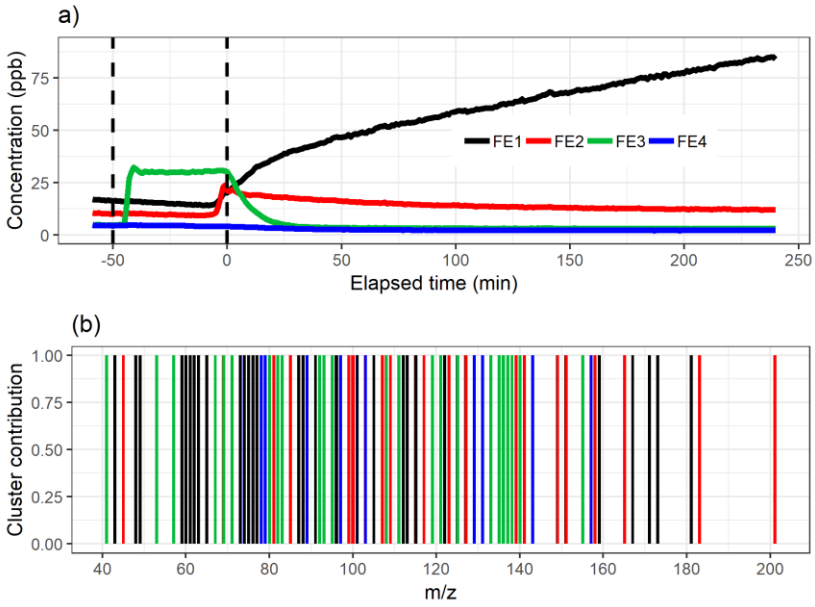

Figure S17 Factor time series (a) and contribution of ion to factor (b) from dichotomized EFA (Oblimin rotated) with 4 factors. The colour code identifying clusters is the same in both panels.

S4.2 NMF and PMF

Figure S18 shows the NMF results with only 4 factors. Figure S19 shows a boxplot of the distribution of the residuals for NMF with 4- and 5-factor solutions. Figure S20 shows the PMF results with factorization rank 4 for the different error schemes.

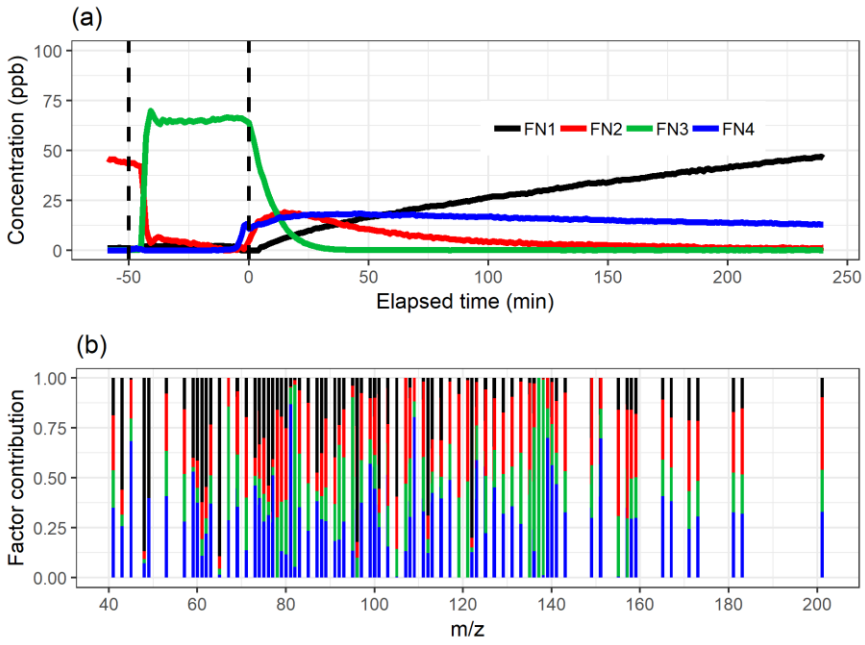

Figure S18 Factor time series (a) and contribution of ion to factor (b) from NMF with 4 factors. The colour code identifying clusters is the same in both panels.

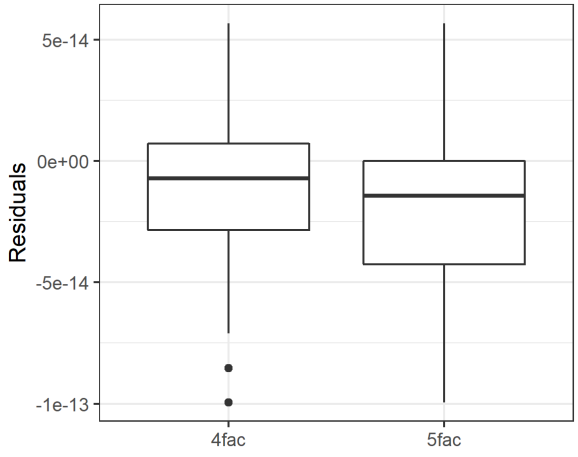

Figure S19 Boxplot of the residuals (original total signal – reconstructed total signal) with 4 and 5 factors from NMF.

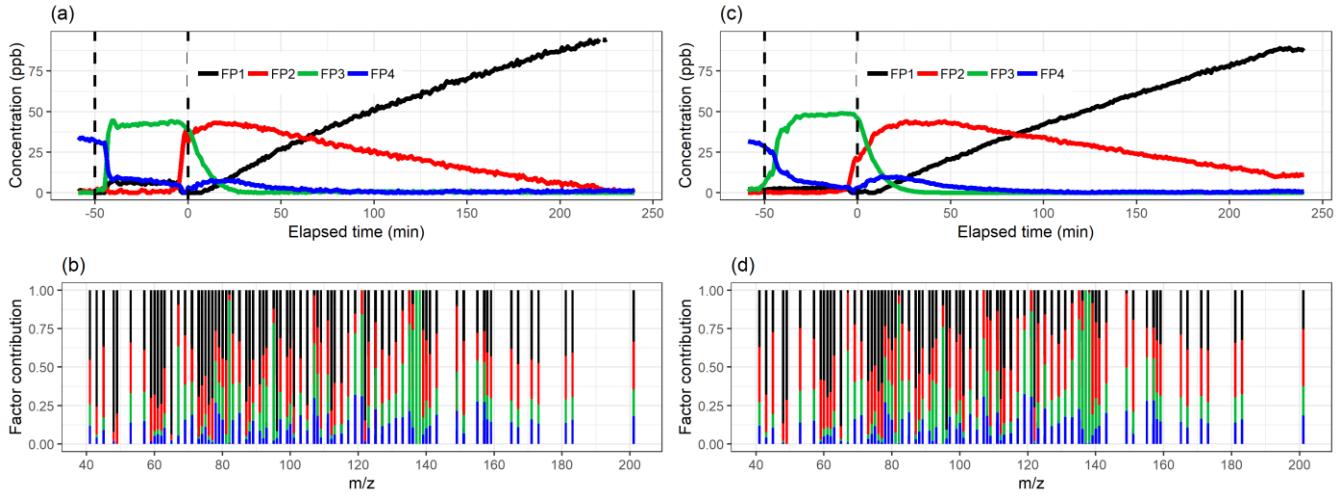

Figure S20 Factor time series and contribution from PMF (fpeak = 0) with static error (a-b) and signal following error (c-d) for factorization rank 4. The colour code identifying the factors is the same in the top and bottom panels.

S5 Contrast angle and factor spectra comparison

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

$$\cos(\theta) = \frac{(\vec{u}\cdot\vec{v})}{(||\vec{u}||\cdot||\vec{v}||)}, \quad \text{(S3)}$$

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 (θ = [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 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

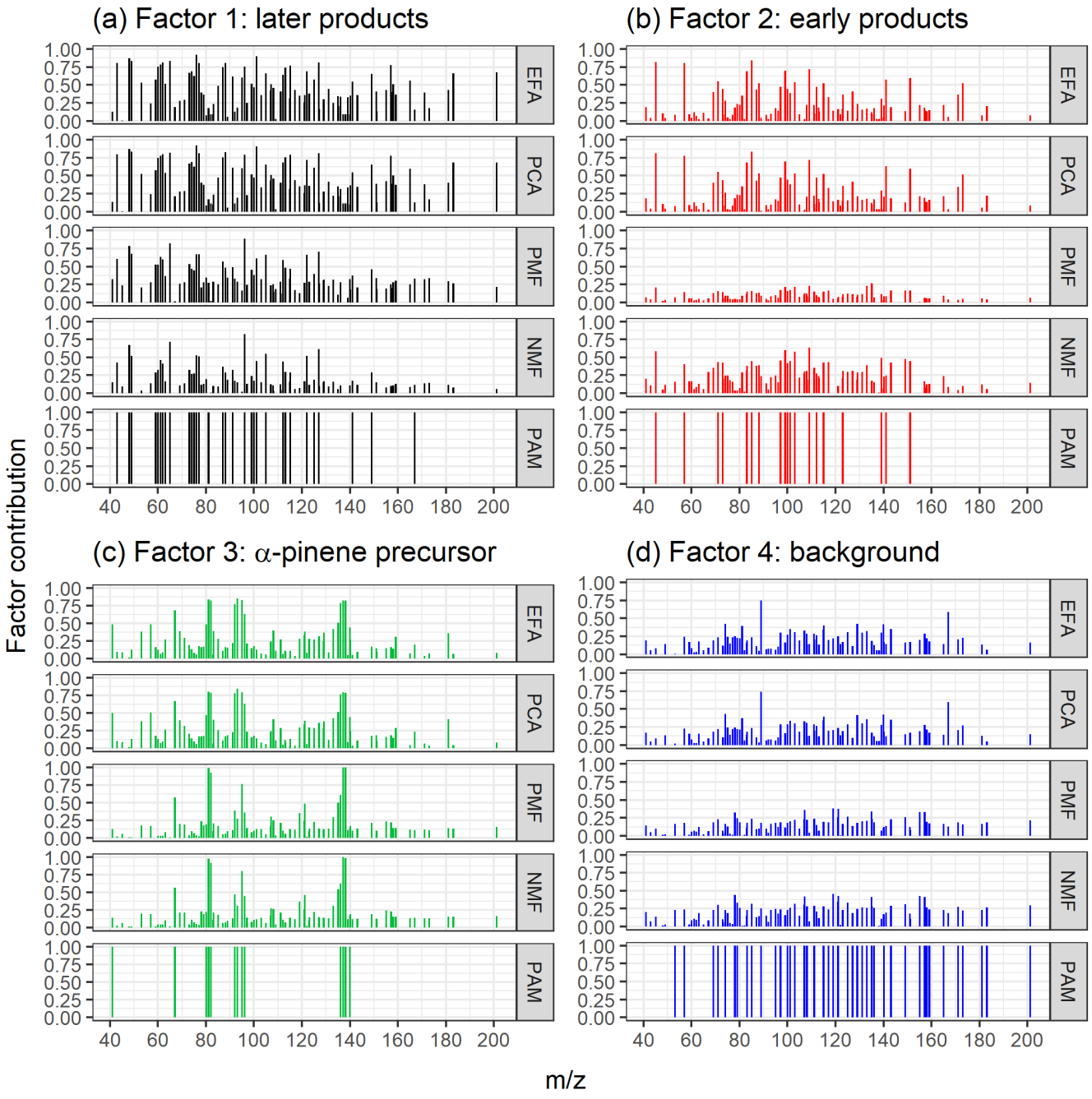

Figure S21 Factor contribution for factors 1-4 from ml-EFA, EVD-PCA, PAM, NMF and PMF with static error for the gas phase data measured with PTR-MS.

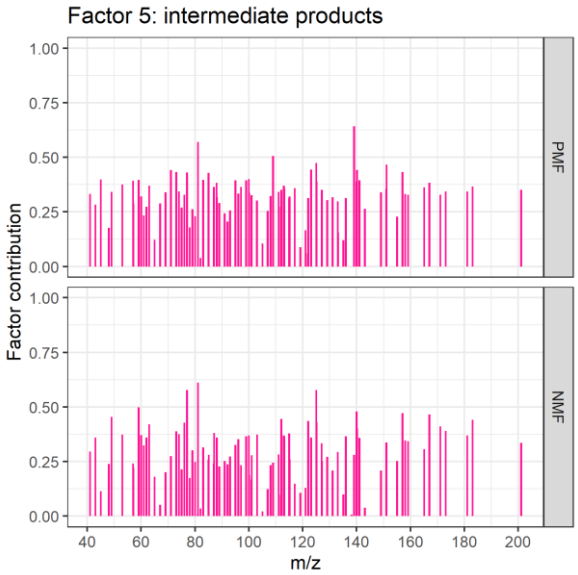

Figure S22 Factor contribution for factor 5 from NMF and PMF with static error for the gas phase data measured with PTR-MS.

S6 Additional SDRT results for AMS data

S6.1 EFA, PCA and PAM

Figure S23 shows the test results for the different number of factors, components and clusters for EFA, PCA and PAM, respectively. Figures S24 and S25 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. S24b barely changes when adding a new factor (Fig. S25b), indicating the algorithm is not able to separate any additional factors from the first 2. Figures S26 and S27 show the results from PAM with 2-5 clusters.

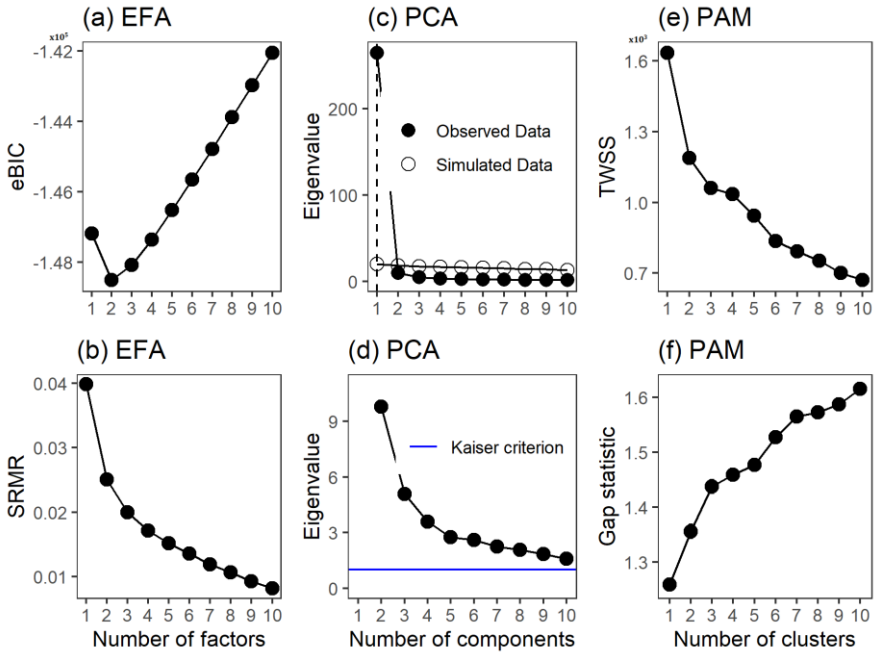

Figure S23 Factor number indexes for particle phase data (AMS). Empirical BIC (a) and SRMR (b) as a function of the number of factors for pa-EFA, Parallel analysis (c) and Kaiser criterion (d) for EVD-PCA and TWSS (e) and Gap statistic (f) for PAM.

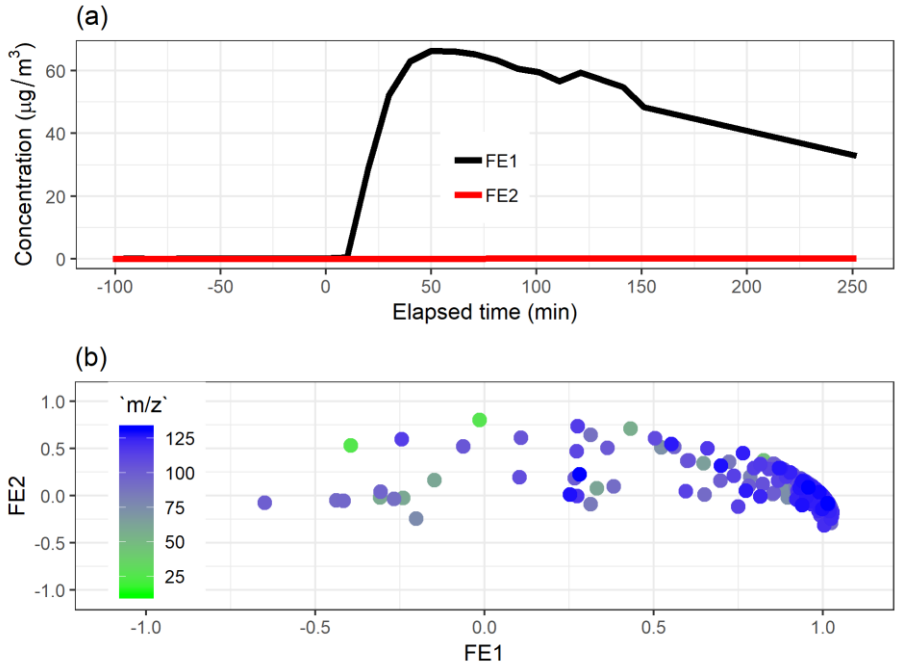

Figure S24 Oblimin rotated factor time series (a) and original loadings (b) from pa-EFA with 2 factors for the particle phase data measured with AMS.

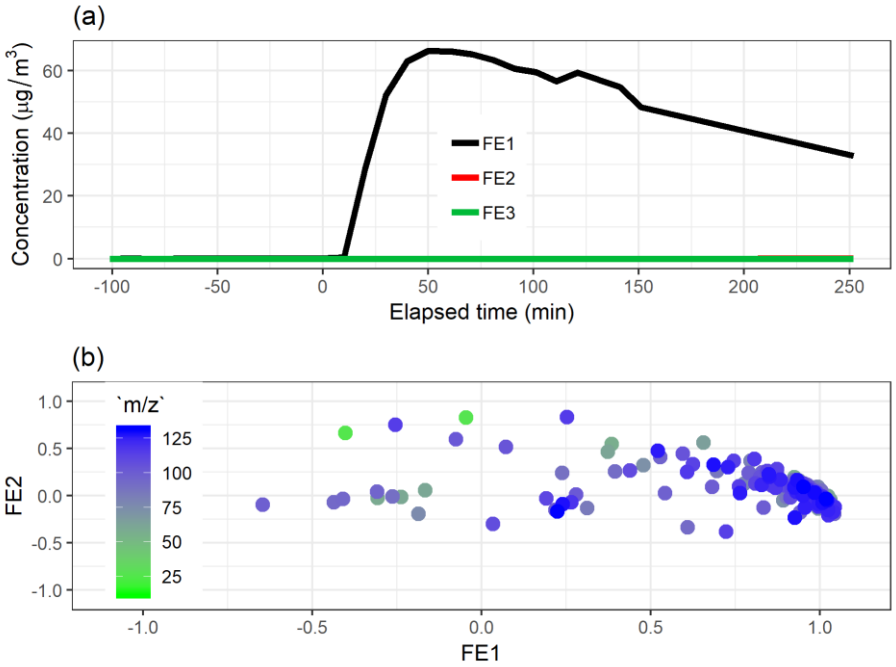

Figure S25 Oblimin rotated factor time series (a) and original loadings (b) from pa-EFA with 3 factors for the particle phase data measured with AMS.

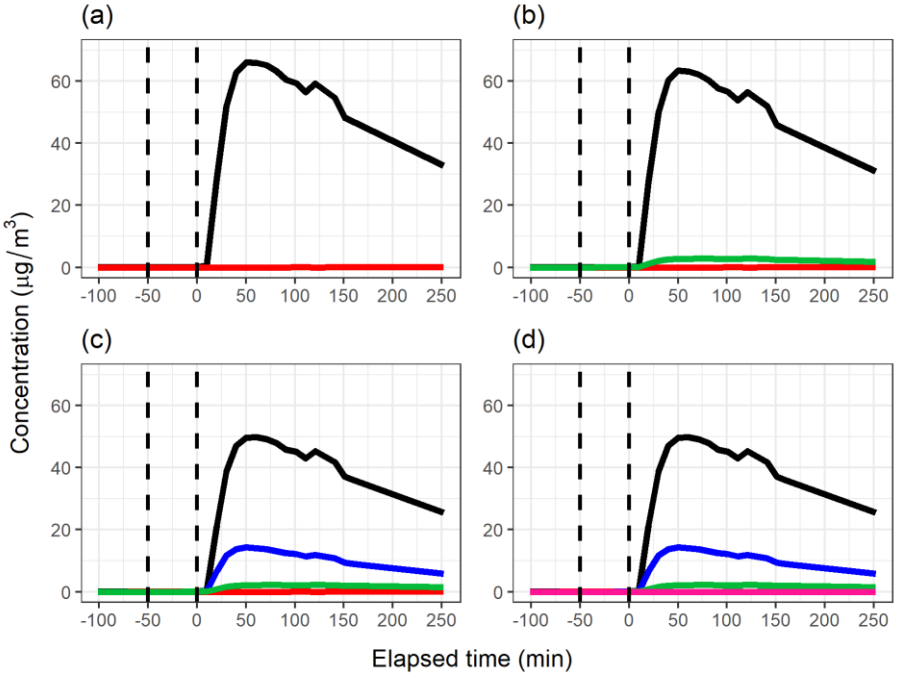

Figure S26 Cluster time series from PAM with (a) 2, (b) 3, (c) 4 and (d) 5 clusters for the measured particle phase (AMS) data.

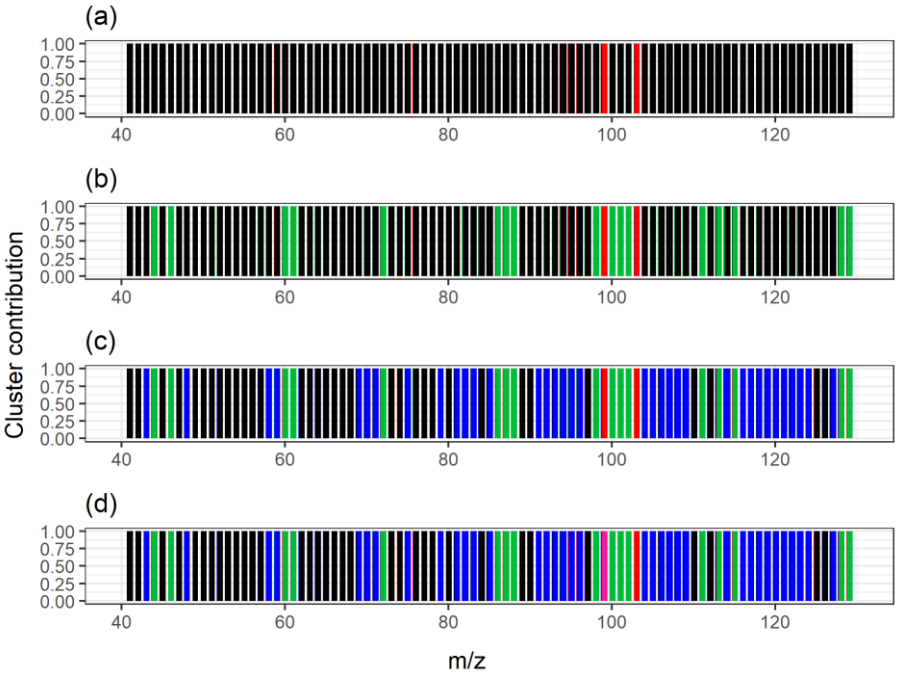

Figure S27 Cluster contribution from PAM with (a) 2, (b) 3, (c) 4 and (d) 5 clusters for the measured particle phase (AMS) data. The colours in the subplots correspond to the time series in Fig. S26. Ions with m/z > 40 are omitted from the plot as no changes were observed (ions were in black factor in all cases).

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. S28), 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. S29) 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.

Figures from S30 to S32 show the PMF results with 3 to5 factors.

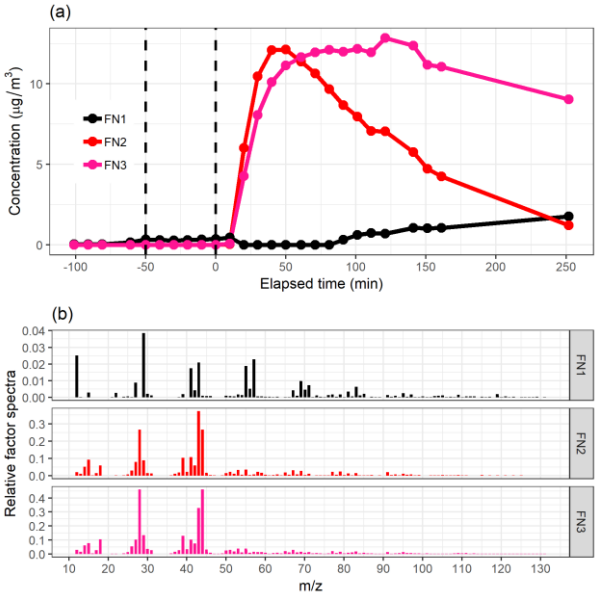

Figure S28 Factor time series (a) and relative factor spectra (b) from NMF with 3 factors for the measured particle phase (AMS) data. The colour code identifying the factors is the same in both panels.

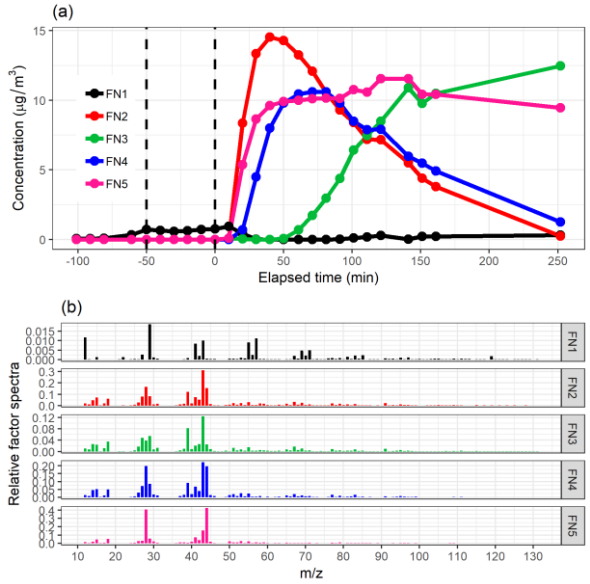

Figure S29 Factor time series (a) and relative factor spectra (b) from NMF with 5 factors for the measured particle phase (AMS) data. The colour code identifying the factors is the same in both panels.

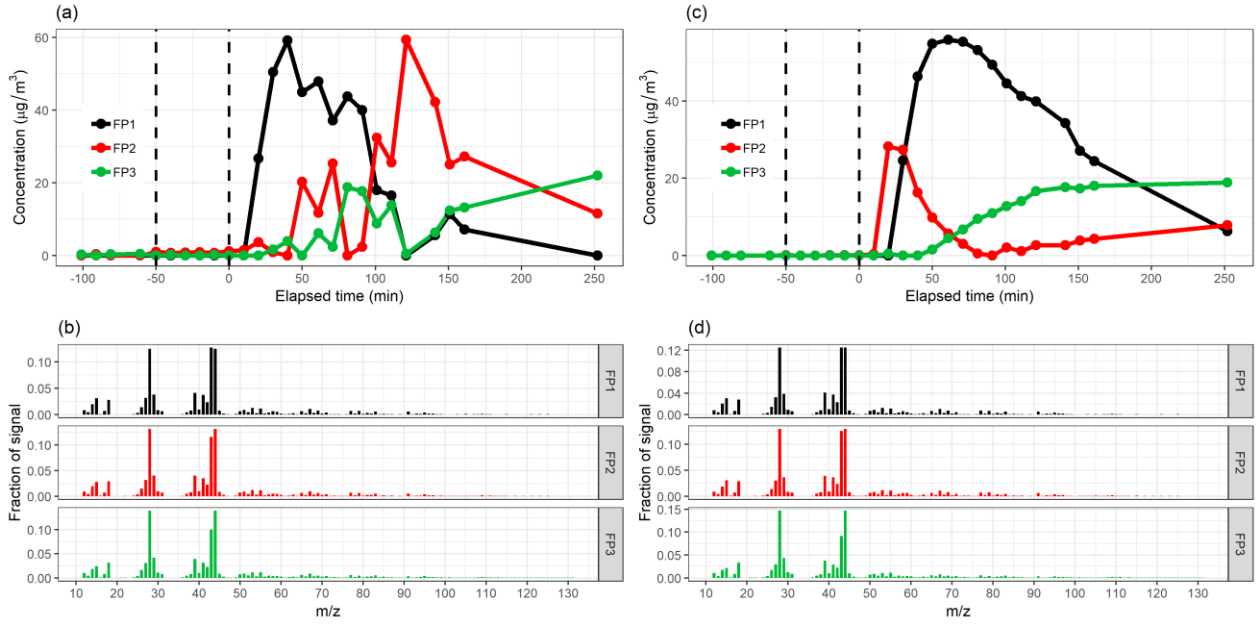

Figure S30 Factor time series and contribution from PMF with static error (a-b) and Standard AMS error (c-d) for factorization rank 3 for the measured particle phase (AMS) data.

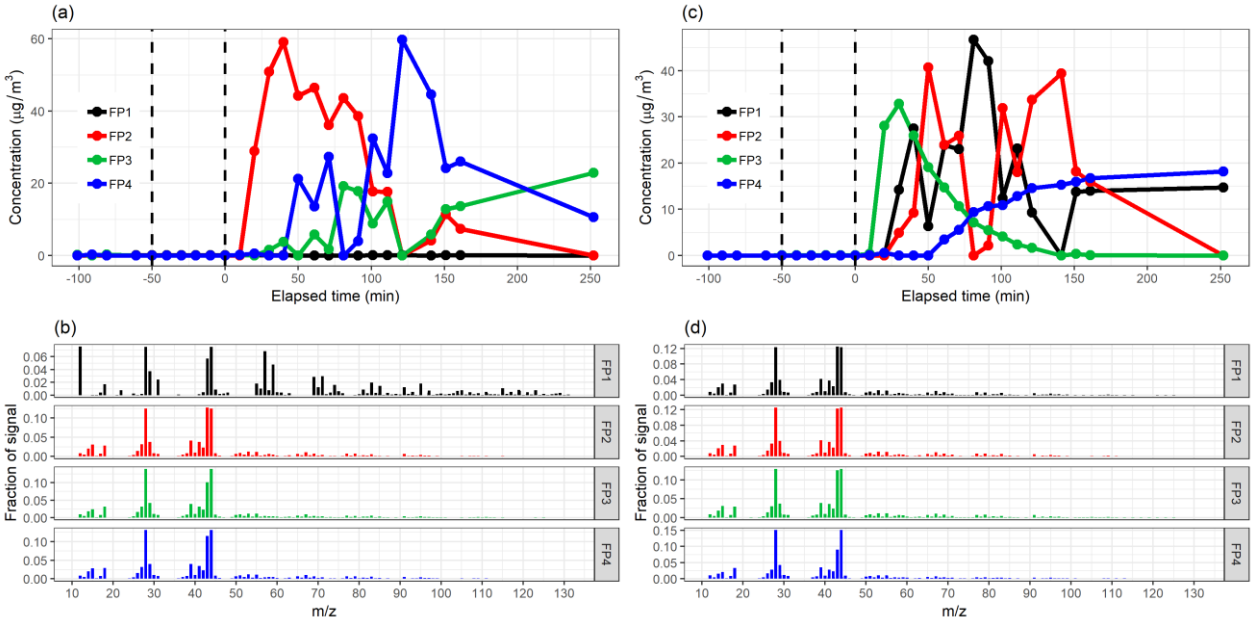

Figure S31 Factor time series and contribution from PMF with static error (a-b) and Standard AMS error (c-d) for factorization rank 4 for the measured particle phase (AMS) data.

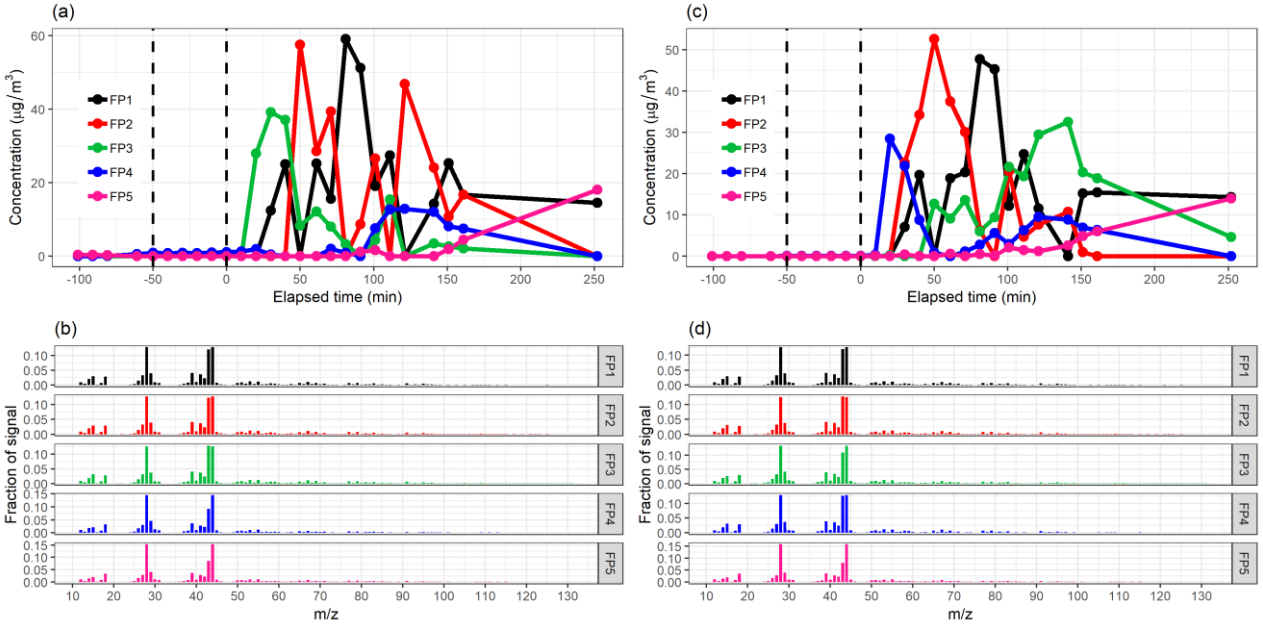

Figure S32 Factor time series and contribution from PMF with static error (a-b) and Standard AMS error (c-d) for factorization rank 5 for the measured particle phase (AMS) data.

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.

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.

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.