Positive Matrix Factorization (PMF) Overview

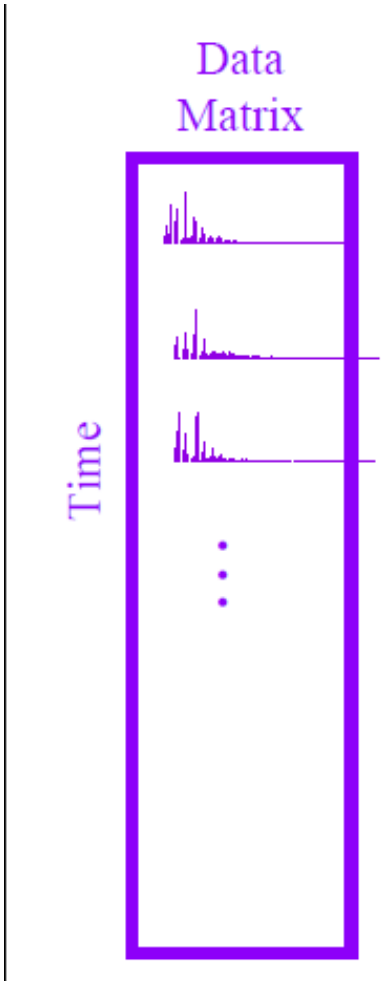
Manjula Canagaratna

AMS/ACSM User Meeting (Virtual)

1/22/2021

Slide Acknowledgements: Ingrid Ulbrich

Overview



AMS/ACSM Mass Spectra have a lot of information.

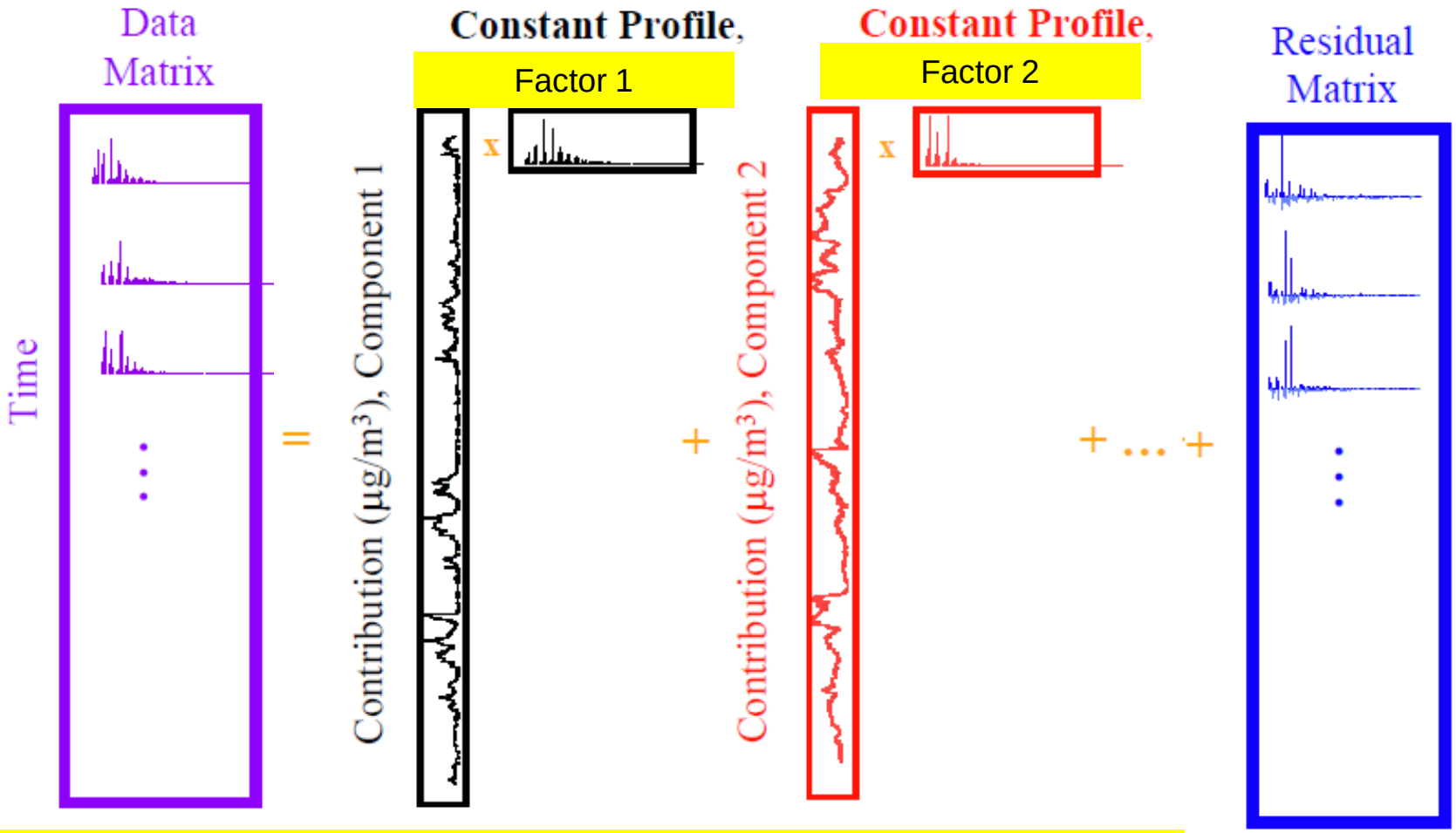
- How are ions correlated in time?
- What are the distinct aerosol sources/process in measured dataset?

Use Positive Matrix Factorization (PMF) to answer these questions

PMF Model

Paatero, P. , 1997

$$X_{ij} = \sum_p G_{ip} F_{pj} + E_{ij}$$



Bilinear Model

Assumptions: Factors profiles (MS) are constant, Factor contributions (TS) vary

PMF

Paatero, P. , 1997

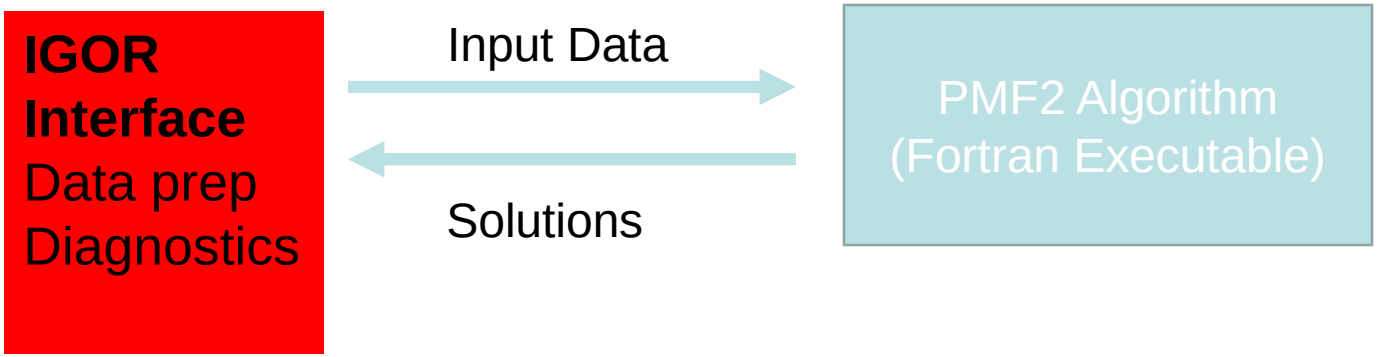
PMF2 Algorithm

Computational method used to solve PMF model

- Weighted Least Squares Fitting in Multiple Dimensions
- Weighting factor is uncertainty in measurements
- Forces non-negative, physically reasonable solutions (i.e. Positive elements for MS and time series)
- User Chooses # of Factors
- PMF- no a-priori knowledge needed

Note: ME-2 algorithm allows solving PMF model w/ apriori information (MS, TS)

PMF Software



IGOR Programs:

1) PET – PMF Evaluation tool (Ulbrich et al., ACP, 2009)
Uses only PMF2 Algorithm to solve PMF model
Freely available (cite source)

2) SOFI – Source Finder (Canonaco et al., AMT, 2013)
Uses ME2 Algorithm to solve PMF and ME2 models
Basic version freely available
(Co-authorship + cite source)
Advanced versions paid subscriptions

PMF Software and interfaces

PMF Model $X_{ij} = \sum_p G_{ip}F_{pj} + E_{ij}$

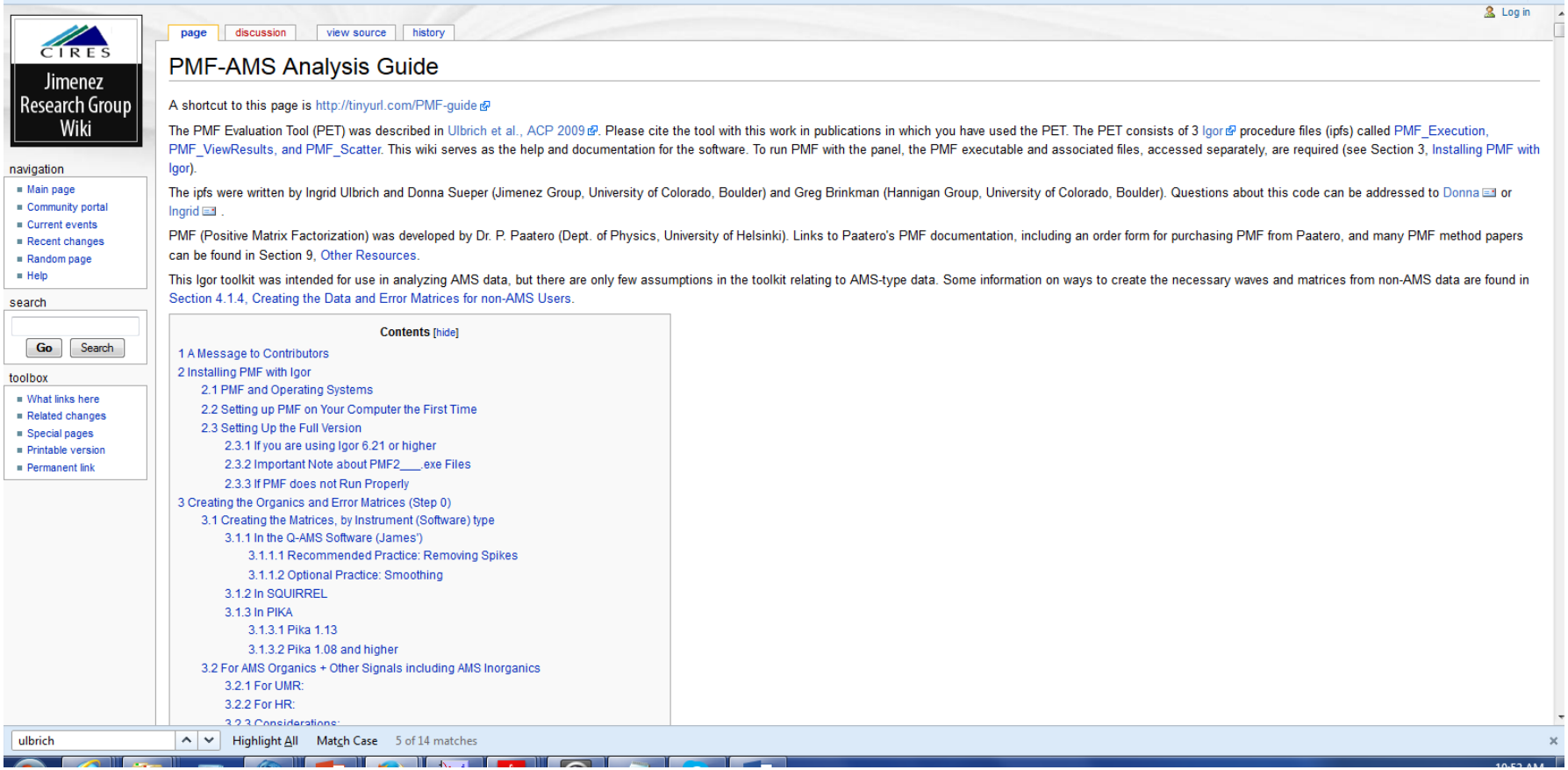
Fortran executables(*.exe) available for purchase from Pentti Paatero for solving this problem

1) PMF2 Algorithm (PMF2.exe, PMF2wopt.exe) – no *apriori* information

2) ME2 Algorithm (ME2.exe) – Allows for *apriori* information about MS and TS, flexible for solving many models besides PMF, slower than PMF2

http://cires1.colorado.edu/jimenez-group/wiki/index.php?title=PMF-AMS_Analysis_Guide

PET (Ulbrich et al. ACP 2009)



CIRES Jimenez Research Group Wiki

page | discussion | view source | history | Log in

navigation
- Main page
- Community portal
- Current events
- Recent changes
- Random page
- Help

search

Go | Search

toolbox
- What links here
- Related changes
- Special pages
- Printable version
- Permanent link

PMF-AMS Analysis Guide

A shortcut to this page is http://tinyurl.com/PMF-guide

The PMF Evaluation Tool (PET) was described in Ulbrich et al., ACP 2009. Please cite the tool with this work in publications in which you have used the PET. The PET consists of 3 Igor procedure files (ipfs) called PMF_Execution, PMF_ViewResults, and PMF_Scatter. This wiki serves as the help and documentation for the software. To run PMF with the panel, the PMF executable and associated files, accessed separately, are required (see Section 3, Installing PMF with Igor).

The ipfs were written by Ingrid Ulbrich and Donna Sueper (Jimenez Group, University of Colorado, Boulder) and Greg Brinkman (Hannigan Group, University of Colorado, Boulder). Questions about this code can be addressed to Donna or Ingrid .

PMF (Positive Matrix Factorization) was developed by Dr. P. Paatero (Dept. of Physics, University of Helsinki). Links to Paatero's PMF documentation, including an order form for purchasing PMF from Paatero, and many PMF method papers can be found in Section 9, Other Resources.

This Igor toolkit was intended for use in analyzing AMS data, but there are only few assumptions in the toolkit relating to AMS-type data. Some information on ways to create the necessary waves and matrices from non-AMS data are found in Section 4.1.4, Creating the Data and Error Matrices for non-AMS Users.

Contents [hide]

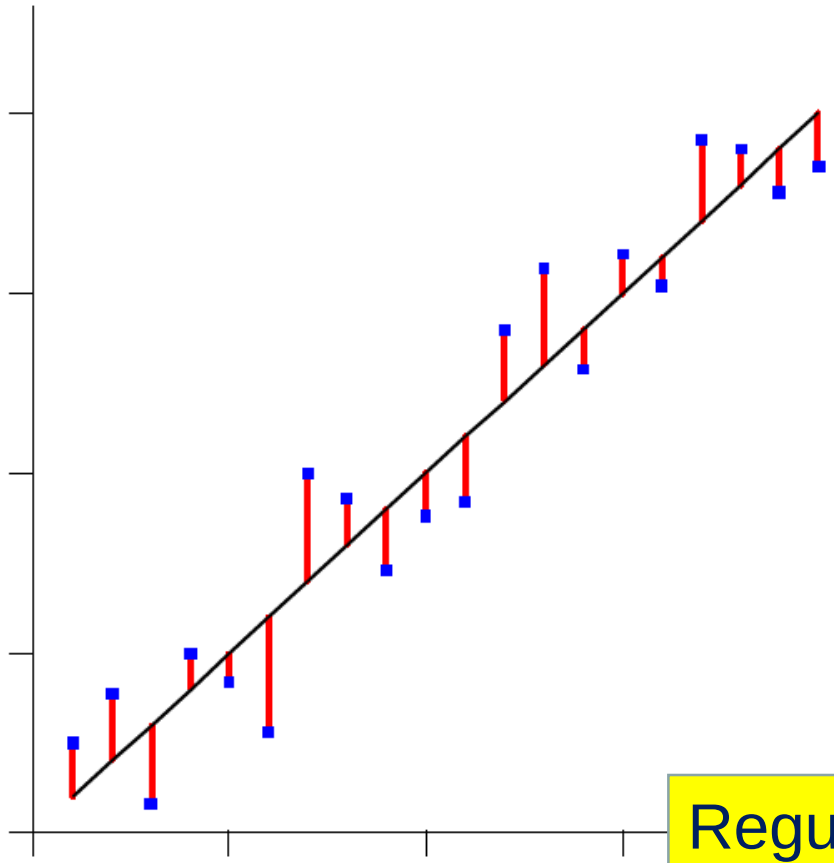
ulbrich | Highlight All | Match Case | 5 of 14 matches

* NOTE: All of the example diagnostics and graphs shown in this presentation are from PET

https://datalystica.com/sofi
SoFi (Canonaco et al. AMT, 2013)

ME-2 allows for using a-priori knowledge to constrain on factor profiles or time trends.

Least-Squares Fit



Have some points…

… and fit a line…

… $\min \sum_i \mathrm{I}^2$

Regular Least-Square minimizes $Chi^2 = \sum_i Resid_i^2$

PMF Approach : Least Squares Fit in multiple dimensions

$$X_{ij} = \sum_p G_{ip} F_{pj} + Resid_{ij}$$

Data matrix

X_{11}	X_{12}	X_{13}	.	.	.
X_{21}					
X_{31}					
.					
.					
.					
.					
.					
.					
.					

Error matrix

σ_{11}	σ_{12}	σ_{13}	.	.	.
σ_{21}					
σ_{31}					
.					
.					
.					
.					
.					
.					
.					

Generated from AMS/ACSM Data matrices

Q matrix

Q_{11}	Q_{12}	Q_{13}	.	.	.
Q_{21}					
Q_{31}					
.					
.					
.					
.					
.					
.					
.					

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{resid_{ij}}{\sigma_{ij}} \right)^2$$

estimated error of each point

Minimized Q during fit

Points with high Q drive the fit (i.e. high residuals relative to σ_{ii})

Data points weighted by σ_{ij}

PMF Approach

- If all points are fit as well as we know them, the Q-contribution from each point is ~ 1

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\sim \sigma_{ij}}{\sigma_{ij}}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\sim 1)^2$$

$$\boxed{Q = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{resid_{ij}}{\sigma_{ij}}\right)^2}$$

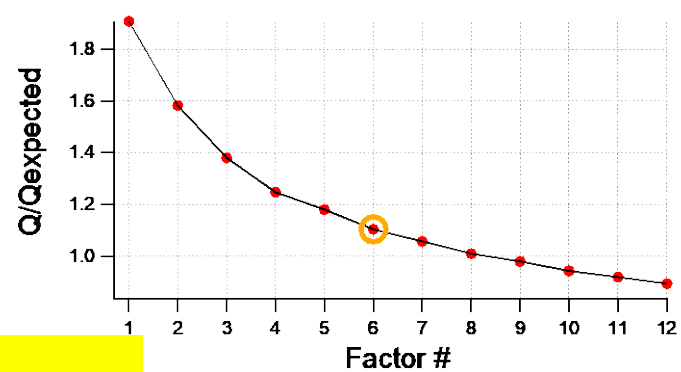
estimated error of each point

and the total Q is ~ the size of the matrix.

For Good Fit, $Q_{exp} \approx m \times n$

Let's normalize the total Q to Q_{exp}, then if all points are fit to $\sim\sigma$, $Q/Q_{exp} \sim 1$

$$Q/Q_{exp} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\sim resid_{ij}}{\sigma_{ij}}\right)^2 / m \times n$$



For Good Fit, $Q/Q_{exp} \approx 1$

Noise Estimates for PMF

- Poisson Statistics (Counting Noise)

$$\Delta I_d = \sqrt{\Delta I_o^2 + \Delta I_b^2} = \alpha \frac{\sqrt{I_o + I_b}}{\sqrt{t_s}}$$

ΔI_d = Noise in diff signal (Hz)
I_o = open signal (Hz)
I_b = closed signal (Hz)

Set a minimum error that corresponds to 1 count / sample time

Previous AMS work by Allan et al. $\alpha = 1.2$

- Electronic noise(decreases with sqrt(t_s))
- Baseline subtraction

<u>Not included</u>

- m/z calibration shifts, peak width uncertainties
- peak fitting errors (Effect of neighbor intensity and spectral distance on fitting)

Downweighting and Discarding

$$X_{ij} = \sum_p G_{ip} F_{pj} + Resid_{ij}$$

Q matrix

Q_{11}	Q_{12}	Q_{13}	.
Q_{21}			
Q_{31}			
.			
.			
.			
.			
.			
.			
.			

Poor or bad S/N ions

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{resid_{ij}}{\sigma_{ij}} \right)^2$$

estimated error of each point

Error provides weighting of each point in fit

Downweight weak S/N ions by artificially increasing error of those ions in error matrix.

Discard Bad S/N ions (Garbage in, Garbage out!)

Downweighting of ions

1) Poor S/N or Bad S/N (included in error preparation - PET)
   Downweight if poor, **Remove if bad**.

2) Data matrix contains multiple copies of same information
   Qm/z 44 only = Qall m/z 44-related ions (included in error preparation - PET)
   m/z 44
   m/z 28=m/z 44
   m/z 18= 0.225* m/z 44
   m/z 17= 0.27 * m/z 44
   m/z 16=0.04 m/z 44
   There are 5 copies of the same information in Q matrix.

Downweight by increasing the noise by a factor of sqrt (# copies)
See Ulbrich et al., ACP, 2009

3) Time trend of ion is observed to be noisier than the calculated noise. (Determined iteratively after running initial PMF a couple times. Manual changes to the errors by user)

Example: CHO^+ (m/z 29)

- Interference from noisy estimate of N_2 isotope- not included in the estimated noise
- Downweight, or discard depending on how bad the problem.

Do units of Input Matrix (Hz vs Nitrate Equiv ug/m3) affect PMF solution

NO

- Any operation that does not affect the temporal correlation of m/zs does not affect PMF Results.

(i.e. if it scales all columns and rows of the data matrix then noise increases proportionally and relative weighting is maintained)

- Operations that affect scaling of only portions of the matrix affect relative weighting of m/zs or time points with respect to each other and DO affect PMF results

NOTE:

Application of RIE * CE can be applied to time trends of each factor time series appropriately. Typically, the same RIE*CE is applied to all organics, but this provides potential to apply different values for HOA vs. OOA, for example.

Combined PMF of Organic +Inorganic

Can provide useful information for apportioning NO_3 and SO_4 into organic and inorganic components

Make combined input matrix from PMF output files from the appropriate m/zs from data and error matrices of the different organic and inorganic species

1) For High- Resolution

Nitrate: Add NO^+, NO_2^+

Sulfate: SO^+(m/z 48), SO_2^+ (m/z 64), SO_3^+ (m/z 80), HSO_3^+ (m/z 81), and H_2SO_4 (m/z 98)

Ammonium: NH^+(m/z 15), NH_2^+ (m/z 16), NH_3^+ (m/z 17)

Chloride: Cl^+ (m/z 35), HCl^+ (m/z 36) for chloride

Do not include ions that are copies or constrained, i.e., scaled to ion intensities (e.g., isotopic ions), are not included. If you do, they need to be downweighted

2) For UMR

Select key inorganic ions that are not scaled copies of others if possible.

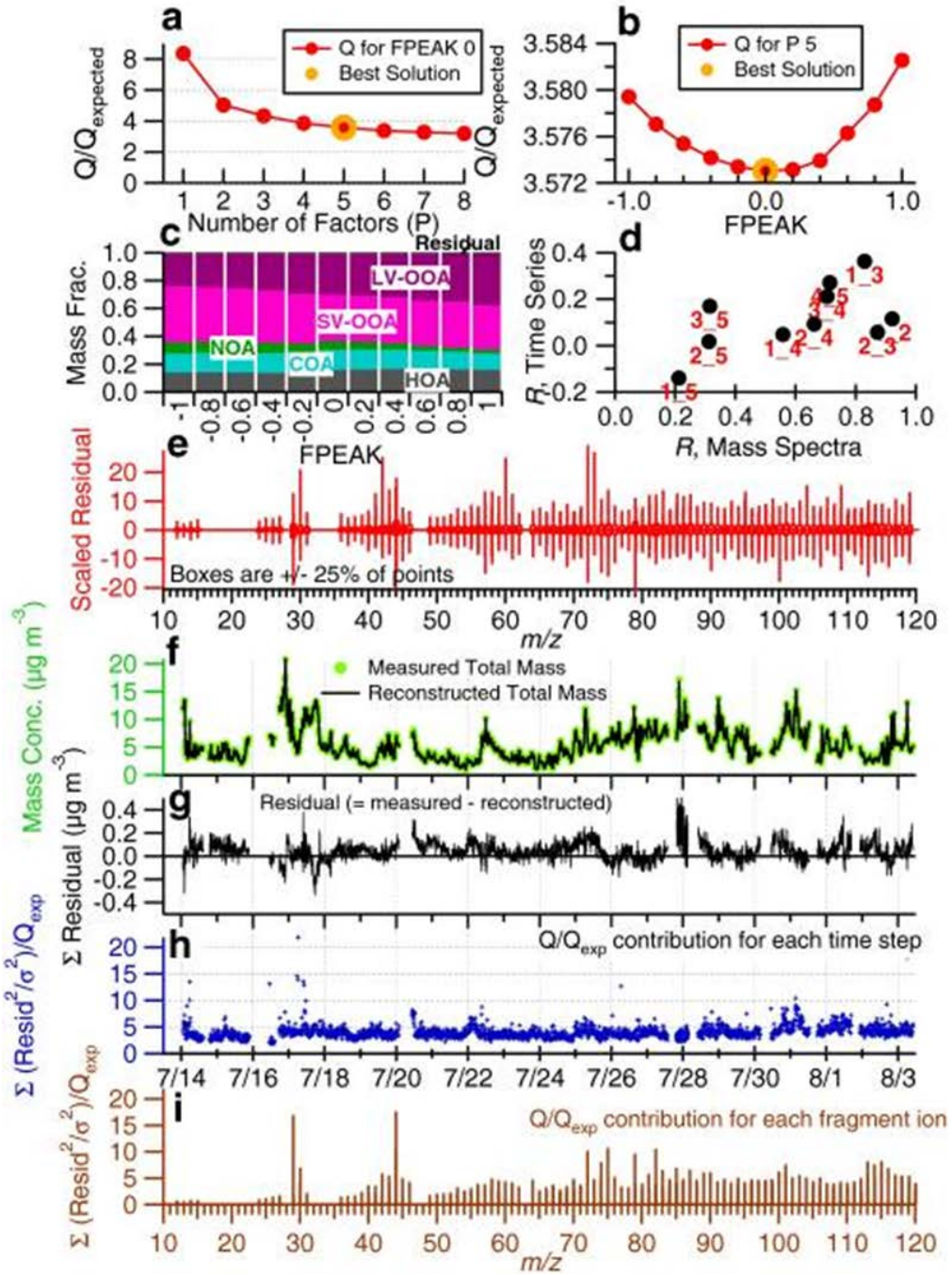
If you want to keep them in to see pattern, Careful of how you deal w/ error matrices

1) Downweight error values for each species according to ions that are duplicated (e.g. m/z 15 (NH^+)=0.1*NH_2^+,m/z 16 (NH_2^+), m/z 17 (NH_3^+)= 1.1*NH^+) . Downweight ions by Sqrt(3)

2) For m/zs that have contributions from multiple species, add errors from each species at m/z in quadrature

PMF Diagnostics to evaluate factor choice

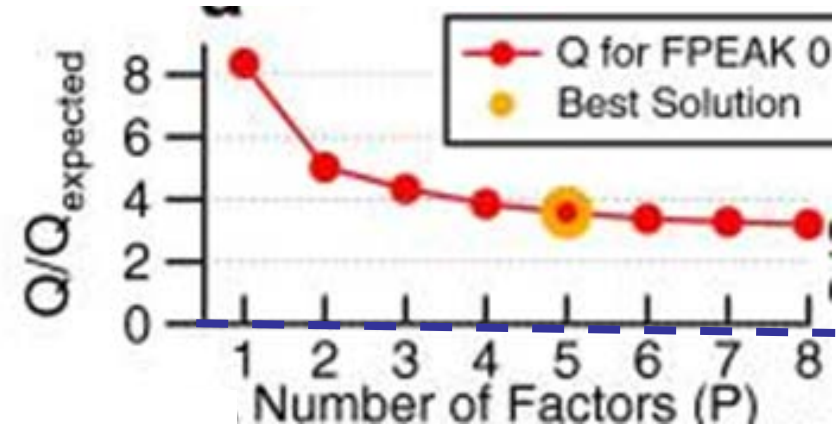
- How well does it reproduce data?
- Is this the "best" solution



Review Article: (Zhang Q. et al., Anal Bioanal Chem, 2011)

PMF Diagnostics

Q/Qexp Distributions



$$\begin{array}{cccccc} Q_{11} & Q_{12} & Q_{13} & . & . & . \\ Q_{21} & & & & & \\ Q_{31} & & & & & \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ . & & & & & \end{array}$$

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{resid_{ij}}{\sigma_{ij}}\right)^2$$

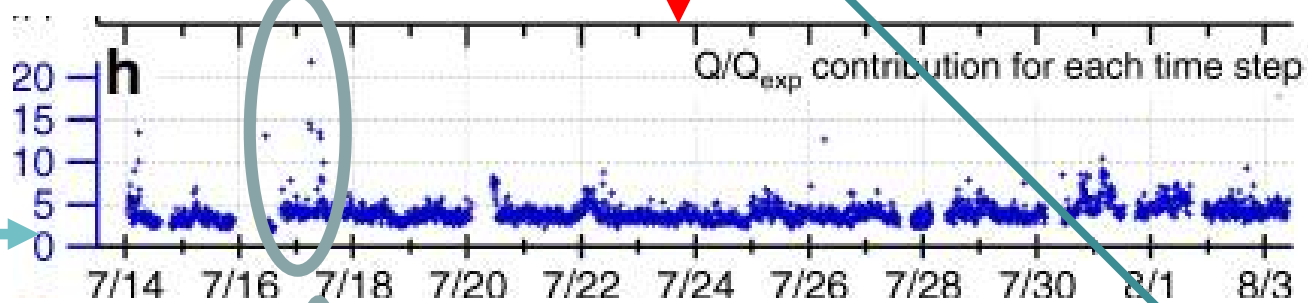
estimated error of each point

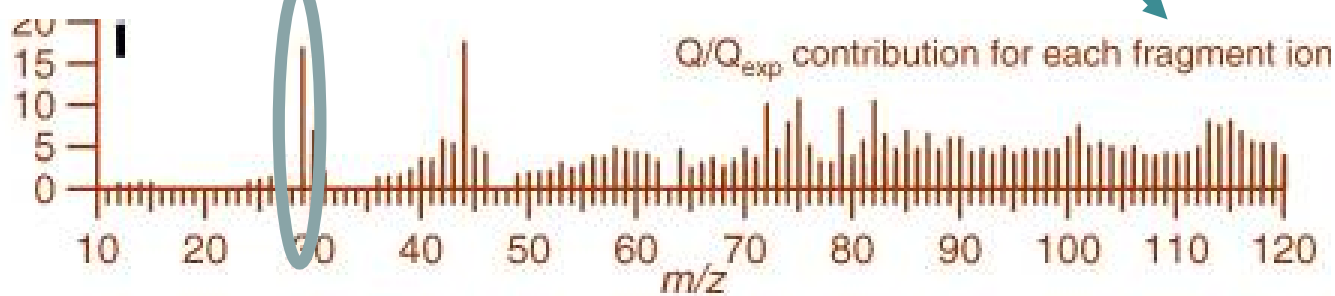
What features are driving the fit?

Large Q/Qexp for some points (or mzs)

Q/Qexp>1 for most of dataset.
-Underestimate Errors?

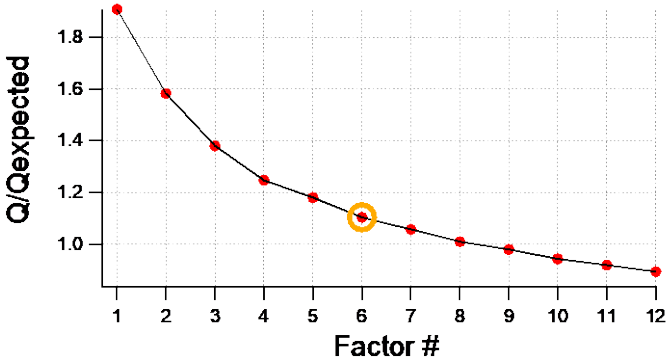
- Dataset doesn't follow PMF model assumptions?





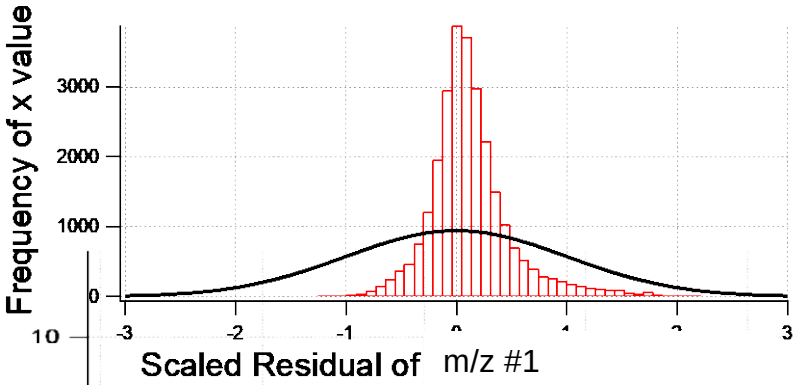
PMF Diagnostics

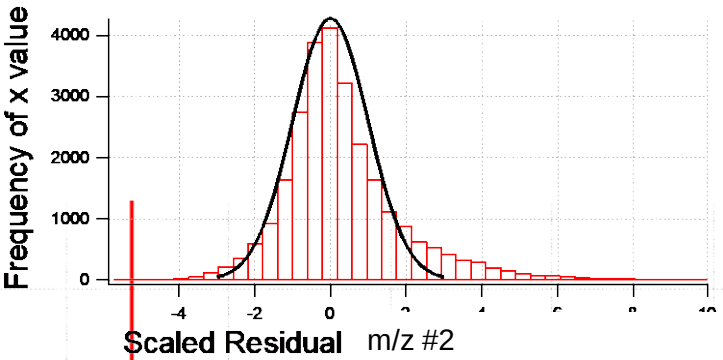
Scaled Residual Distribution

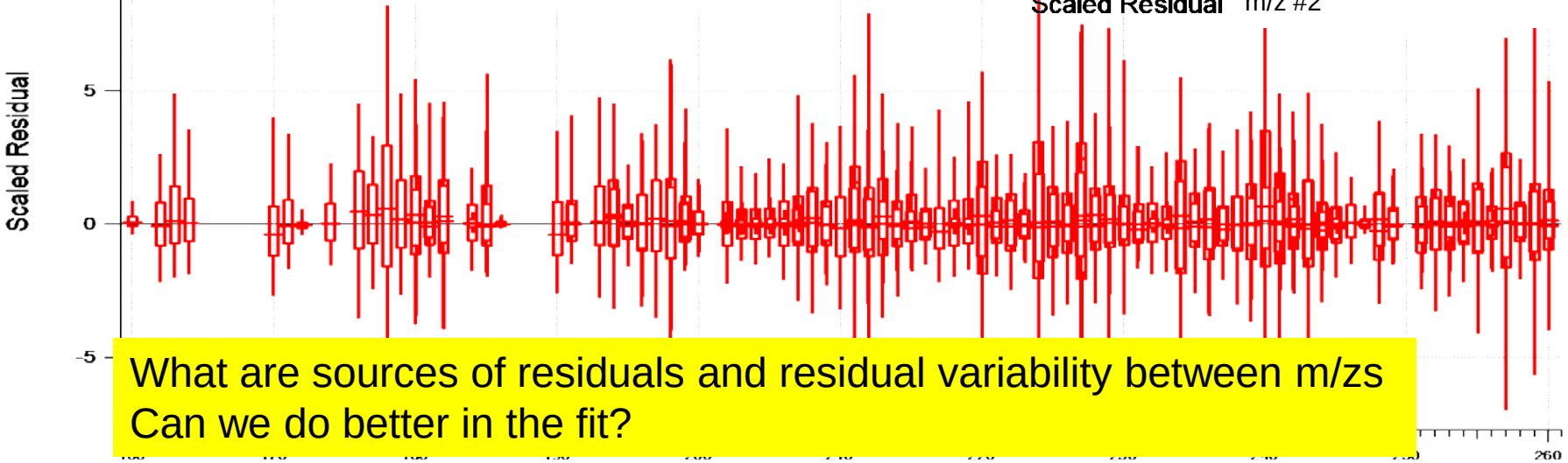


$$X_{ij} = \sum_p G_{ip}F_{pj} + Resid_{ij}$$

$$Scaled\ Residual_{ij} = Resid_{ij}/\sigma_{ij}$$

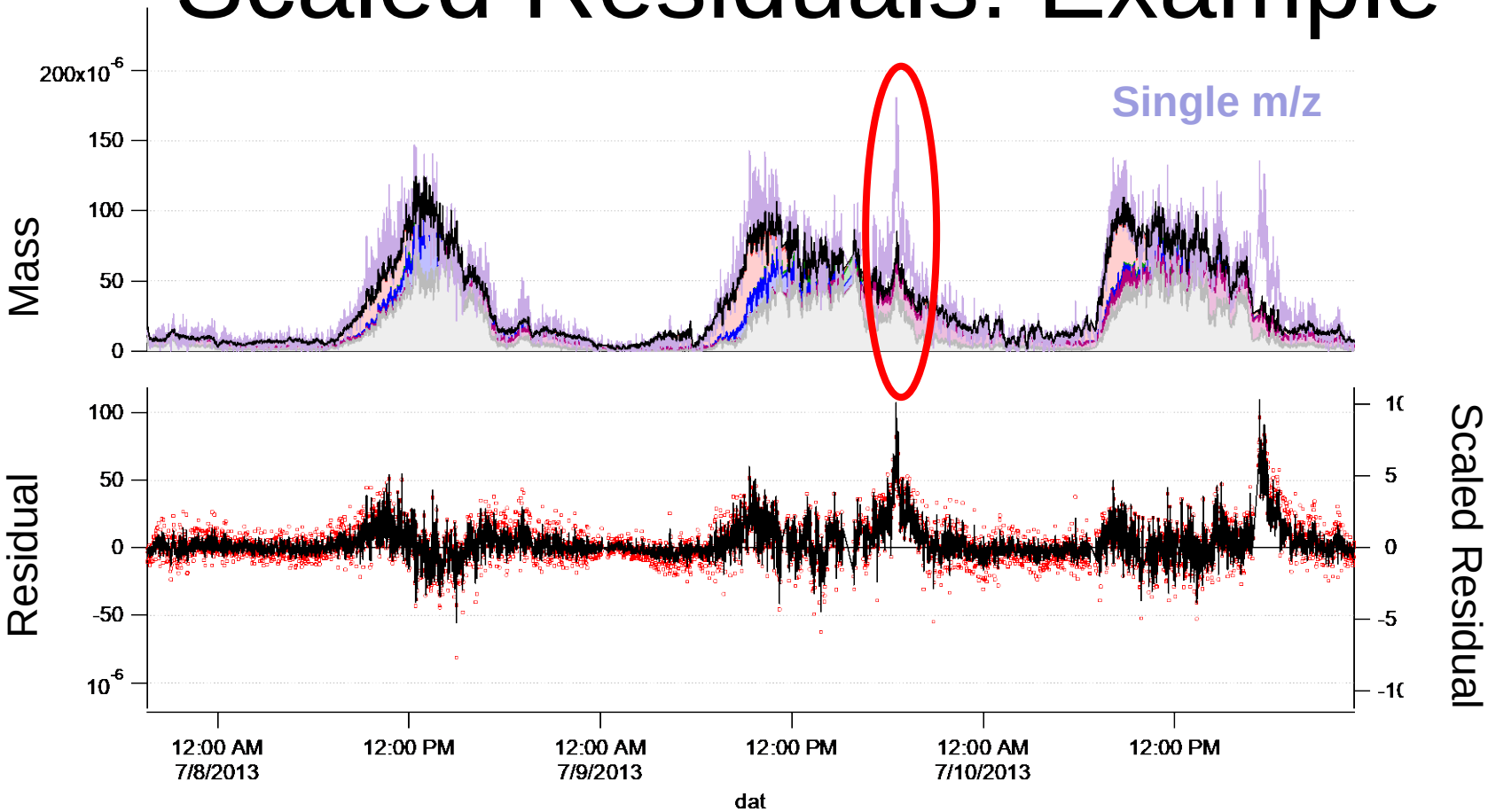






What are sources of residuals and residual variability between m/zs
Can we do better in the fit?

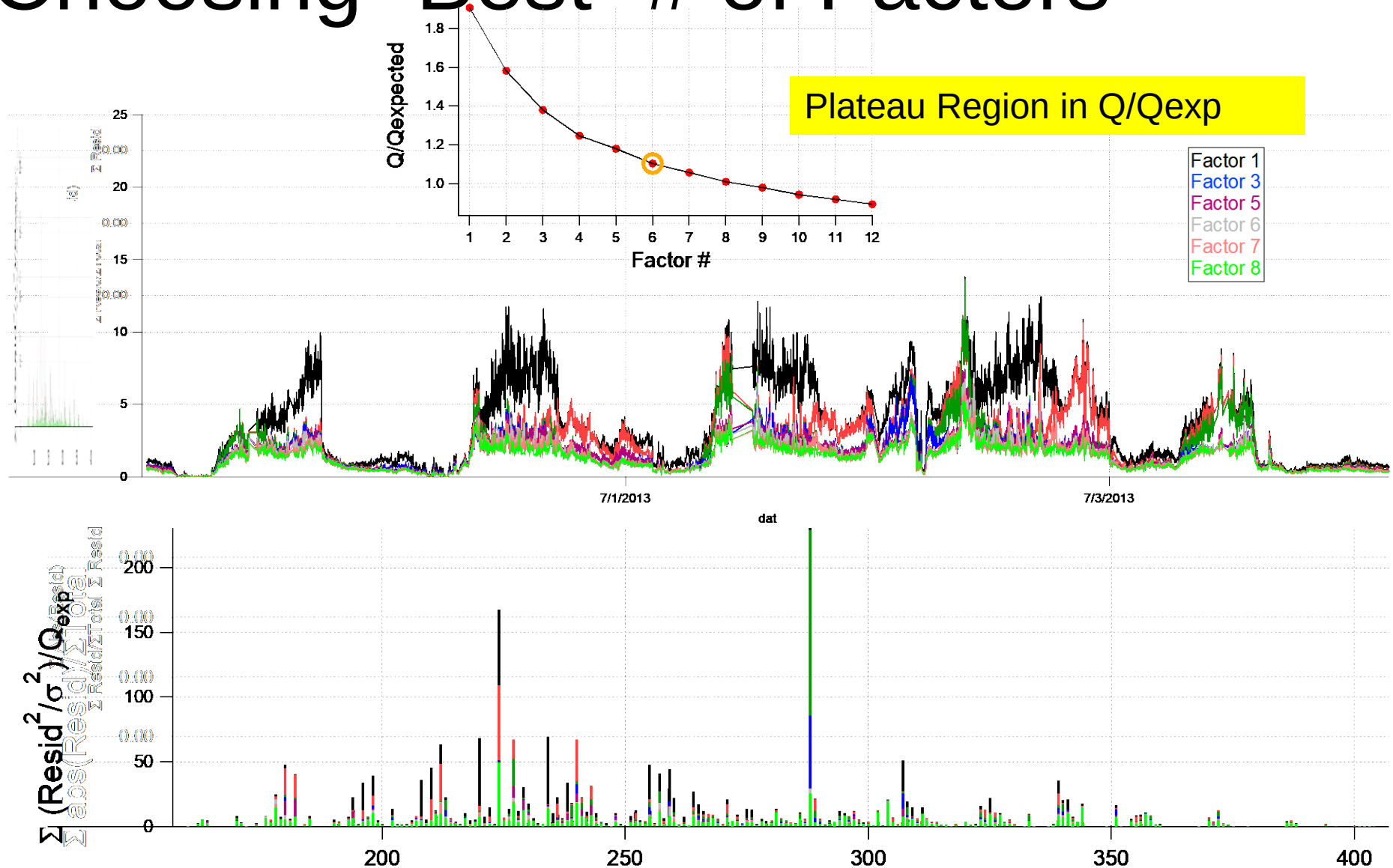
Scaled Residuals: Example



Larger scaled residuals indicate features that are not well described by fit

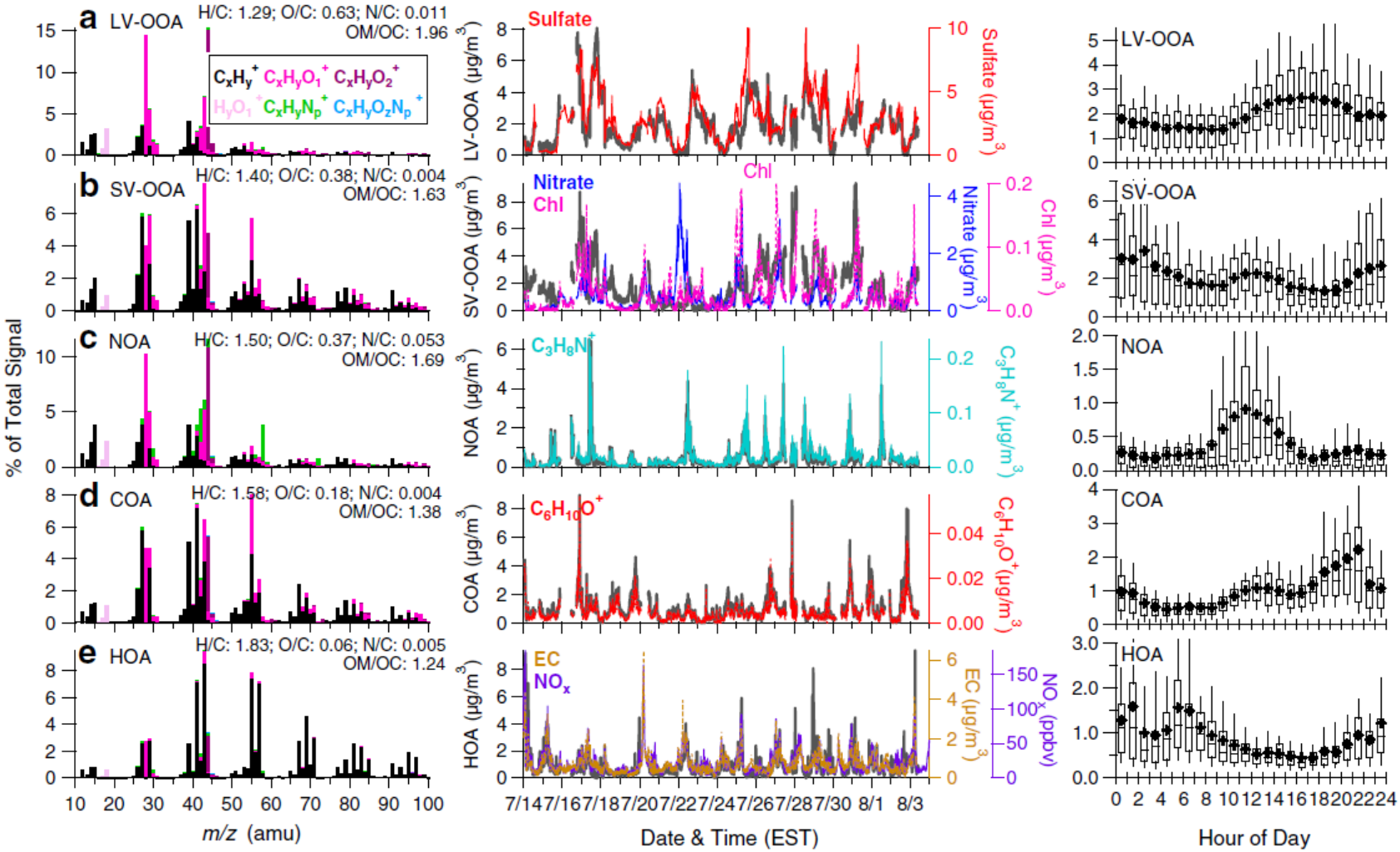
- Addition of factors may improve fit
- May indicate breakdown of constant factor assumption

Choosing “Best” # of Factors



Q/Qexp Distributions as a function of Factor Numbers

Factor Interpretation



Examine Factor MS and compare w/ reference MS

Compare Factor trends with tracer ion and ancillary info

Investigate diurnal trends

Cautions for interpreting PMF

Correlation does NOT mean causation

- Variability in time trends necessary for separation of more factors
- Sources/processes with near identical time-trends and/or spectra cannot be easily resolved with PMF.

→ HR spectra can provide better separation of factors with small mass contributions

→ ME-2 to separate based on small variations

- PMF assumptions of constant factor profiles may not be valid

  Better for HOA than continuum of OOA

Ambiguity in PMF solutions

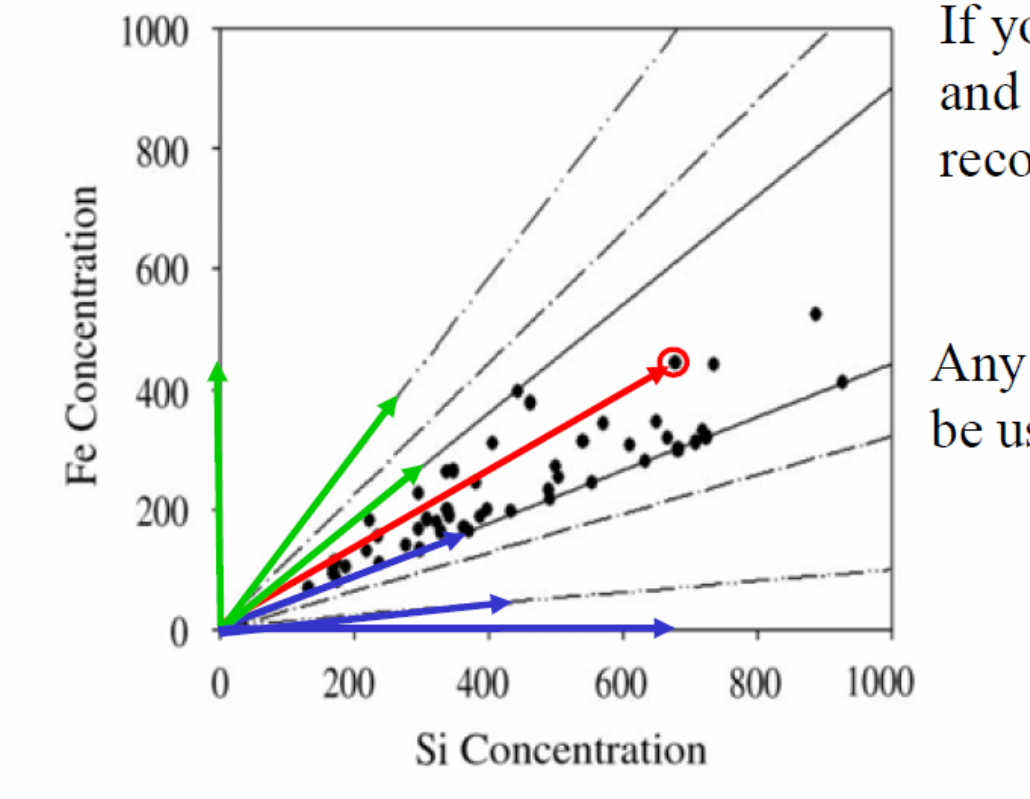


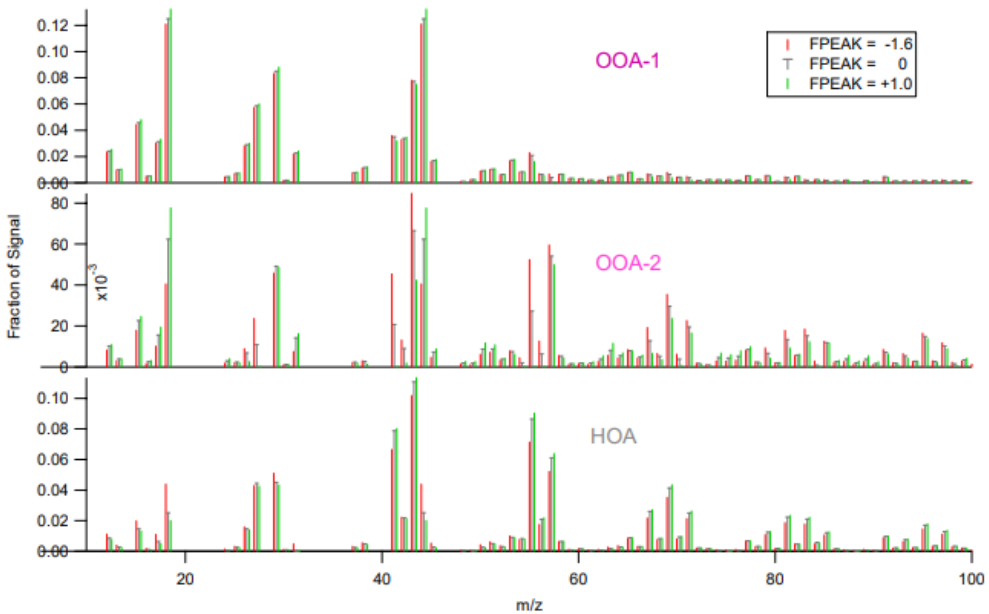
Fig. 1. Simulated data showing multiple possible source "profiles" that could be used to fit the data.

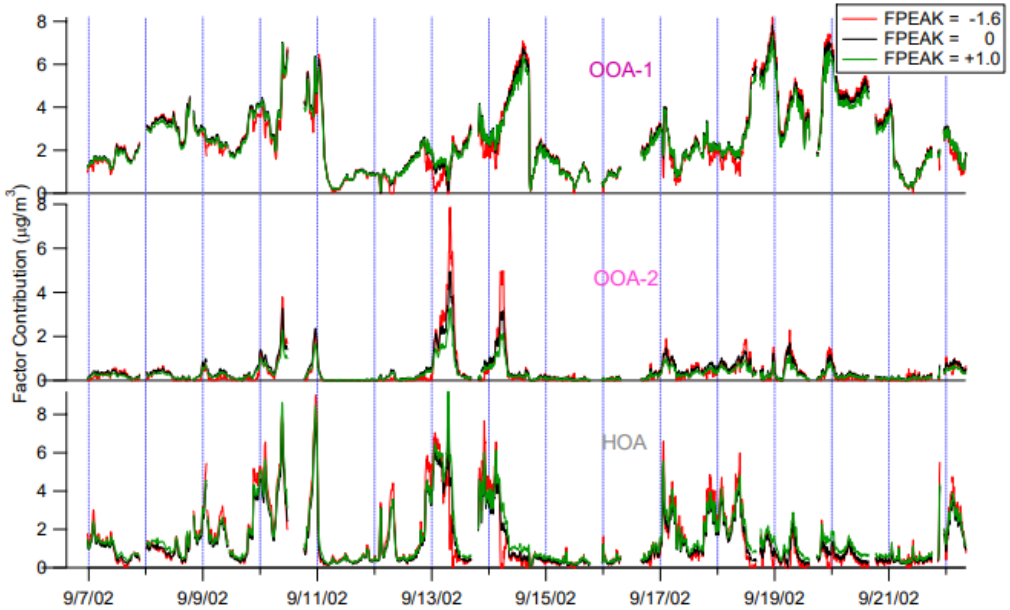
If you only had two species and were trying to reconstruct some data…

Any of these **pairs** could be used.

Paatero et al., Chemometrics and Intelligent Lab Systems, 2002

Rotational ambiguity can make PMF solutions non-unique





Report Uncertainty due to Ambiguity In Factor Mass Spectra and Time Series

- Multiple runs of Fpeak
- Multiple runs of seeds
- Bootstrapping Runs

Ulbrich et al., ACP, 2009

Ways of increasing info from PMF

- Inorganic + organic
- HR + UMR
- Time periods (Long vs. short)

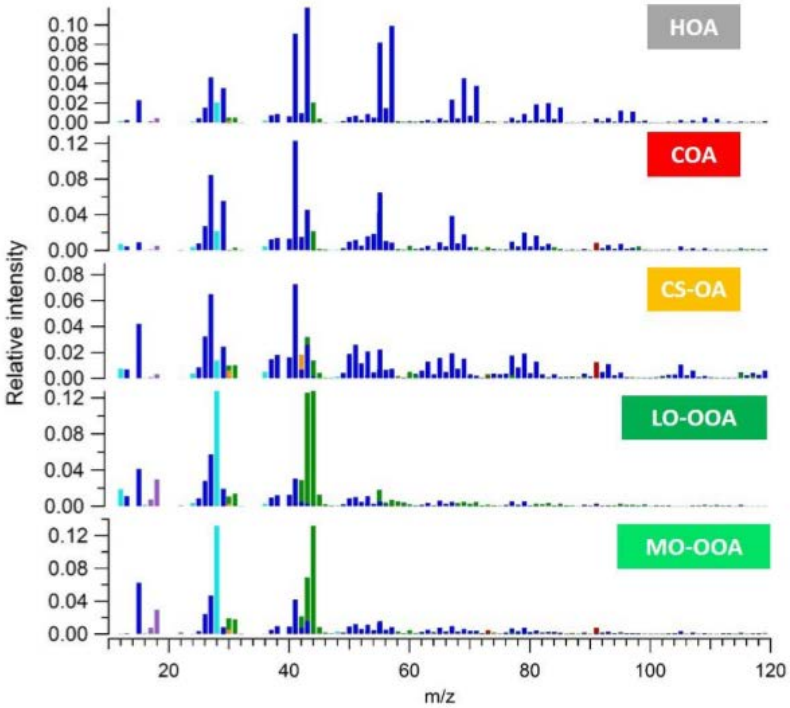
Don’t need to use all the measured ions and all time if they don’t contain variability of interest

Combination w/ other instruments

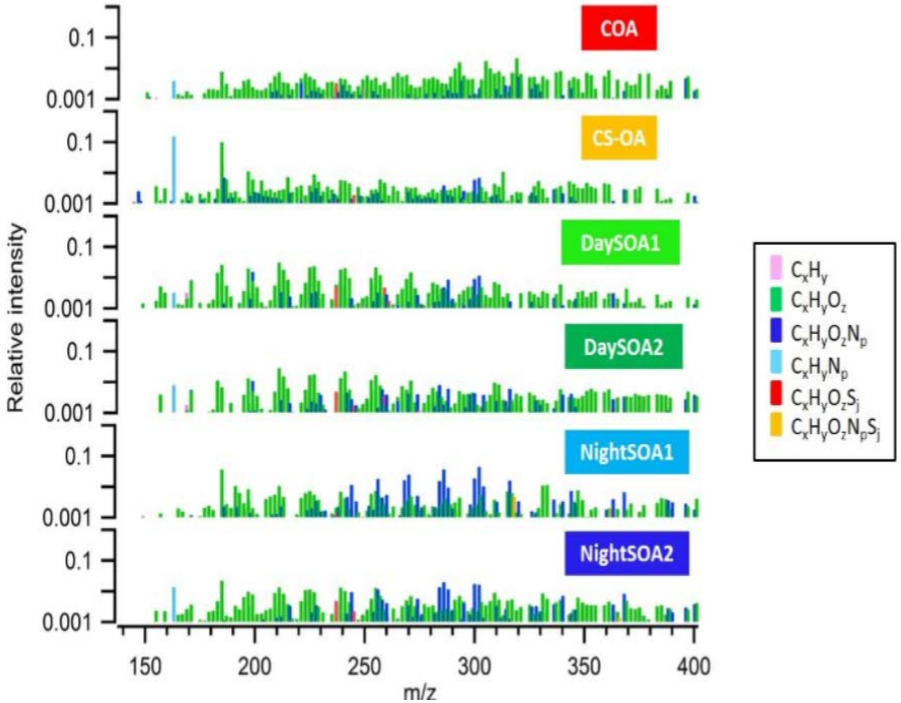
- AMS + PTR
- AMS + EESI (or other particle phase CIMS techniques)

Applications: Source Apportionment

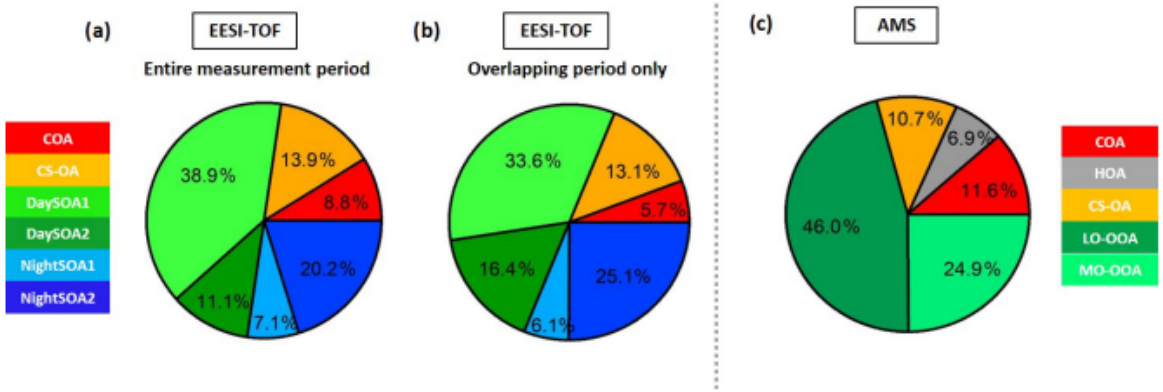
Stefenelli et al., ACP, 2019



AMS



EESI: Increased information about SOA



References

PMF Theory

1) Paatero, P.: Least squares formulation of robust non-negative factor analysis, Chemometr. Intell. Lab., 37, 23–35, https://doi.org/10.1016/S0169-7439(96)00044-5, 1997.

Introduction to Igor software programs

PET:

Ulbrich,I.M.,Canagaratna,M.R.,Zhang,Q.,Worsnop,D.R.,and Jimenez, J. L.: Interpretation of organic components from PositiveMatrixFactorizationofaerosolmassspectrometricdata,Atmos. Chem. Phys., 9, 2891–2918, https://doi.org/10.5194/acp-92891-2009, 2009.

SOFI:

Canonaco, F., Crippa, M., Slowik, J. G., Baltensperger, U., and Prévôt, A. S. H.: SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME2) for the source apportionment: ME-2 application to aerosol mass spectrometer data, Atmos. Meas. Tech., 6, 3649–3661, https://doi.org/10.5194/amt-6-3649-2013, 2013.

References

Other papers describing PMF of AMS data

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Ulbrich, I. M., Ng, N. L., Worsnop, D. R., and Sun, Y.: Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry: a review, Anal. Bioanal. Chem., 401, 3045–3067, https://doi.org/10.1007/s00216-011-5355-y, 2011.

Lanz, V.A., M.R. Alfarra, U. Baltensperger, B. Buchmann, C. Hueglin, and A.S.H. Prévôt, Source Apportionment of Submicron Organic Aerosols at an Urban Site by Factor Analytical Modelling of Aerosol Mass Spectra, *Atmospheric Chemistry and Physics*, 7, 1503-1522, 2007

PMF of Mixed AMS/PTRMS dataset

Slowik et al., Atmos. Chem. Phys., 10, 1969–1988, 2010 www.atmos-chem-phys.net/10/1969/2010/

Questions?

Thank You for Joining and Staying till the End!!!