V. SoFi workshop
General information

+

General - Logistics

- **Food**

– Morning coffee break (~10.30)

– Standing lunch (12.30 – 13.30)

– Afternoon coffee break (~15.00)

➔ Social dinner (Wednesday evening)

- **Internet (Free Internet, check whiteboard)**
- **Restrooms (downstairs)**

General - Logistics

- **Information / presentations / data / code**
  - http://indico.psi.ch/event/SoFi2016

- **Current policy with SoFi**
  - Collaboration for 2 publications per institute (F. Canonaco, A. Prevot and PSI people supporting your analysis during the workshop)
  - Cite the SoFi paper in AMT (Canonaco et al 2013)

General – Program

Monday (Pre-Workshop)	
Time	Activity
10.30 – 12.30	➢ *General support* for communication between SoFi and ME-2, HDF option in Igor ➢ *Theory input* on PMF, ME-2, Q-space, robust mode, rotational tools (a-value, fpeak, pulling)
12.30 – 13.30	******Lunch*****
13.30 – 15.00	➢ *Interactive discussion* using ACSM data in SoFi to better visualize the options/features present in SoFi (import raw data, treat data for PMF run, call PMF, import results in igor for SoFi, explore results)
15.00 – 15.30	***Coffee break***
15.30 – 17.00	➢ *Interactive discussion* using ACSM data in SoFi to better visualize the options/features present in SoFi (import raw data, treat data for PMF run, call PMF, import results in igor for SoFi, explore results)

General – Program

Tuesday (Official kick-off)	
Time	Activity
09.00 – 10.30	➢ *Theory input* on rotational ambiguity, criteria-based approach , propagation of statistical uncertainty, AuRo-SoFi
10.30 – 11.00	***Coffee break***
11.00 – 12.30	➢ *Theory input* on rotational ambiguity, criteria-based approach , propagation of statistical uncertainty, AuRo-SoFi ➢ *Practical example:* Application of SoFi on year-long ACSM data
12.30 – 13.30	******Lunch*****
13.30 – 15.00	➢ *Group discussions:* Users treating similar data, e.g. filter-based, offline, UMR-AMS, HR-AMS, combined datasets have the possibility to share gained experience ➢ *Individual work:* participants work on their own data (support provided)
15.00 – 15.30	***Coffee break***
15.30 – 17.00	➢ *Group discussions:* Users treating similar data, e.g. filter-based, offline, UMR-AMS, HR-AMS, combined datasets have the possibility to share gained experience ➢ *Individual work:* participants work on their own data (support provided)

General – Program

Wednesday	
Time	Activity
09.00 – 10.30	➢ *Individual work:* participants work on their own data (support provided)
10.30 – 11.00	***Coffee break***
11.00 – 12.30	➢ *Individual work:* participants work on their own data (support provided)
12.30 – 13.30	******Lunch*****
13.30 – 15.00	➢ *Presentations of case studies:* source apportionment (SA) studies conducted with SoFi from experienced users (PSI and non-PSI)
15.00 – 15.30	***Coffee break***
15.30 – 17.00	➢ *Presentations* of case studies: source apportionment (SA) studies conducted with SoFi from experienced users (PSI and non-PSI)
***********Social dinner**********	

General – Program

Thursday	
Time	Activity
09.00 – 10.30	➢ *Individual work:* participants work on their own data (support provided) ➢ *Presentations of participants*
10.30 – 11.00	***Coffee break***
11.00 – 12.30	➢ *Presentations of participants* **Conclusion of SoFi workshop**
12.30 – 13.30	******Lunch*****
13.30 – 15.00	**Start of ACTRIS meeting** ➢ *ACTRIS-related discussions*
15.00 – 15.30	***Coffee break***
15.30 – 17.00	➢ *ACTRIS-related discussions*

Friday	
Time	Activity
09.00 – 17.30	➢ *If wished/needed, further discussion at PSI with PSI people at PSI (please announce this during the workshop)*

V. SoFi workshop
PMF - general

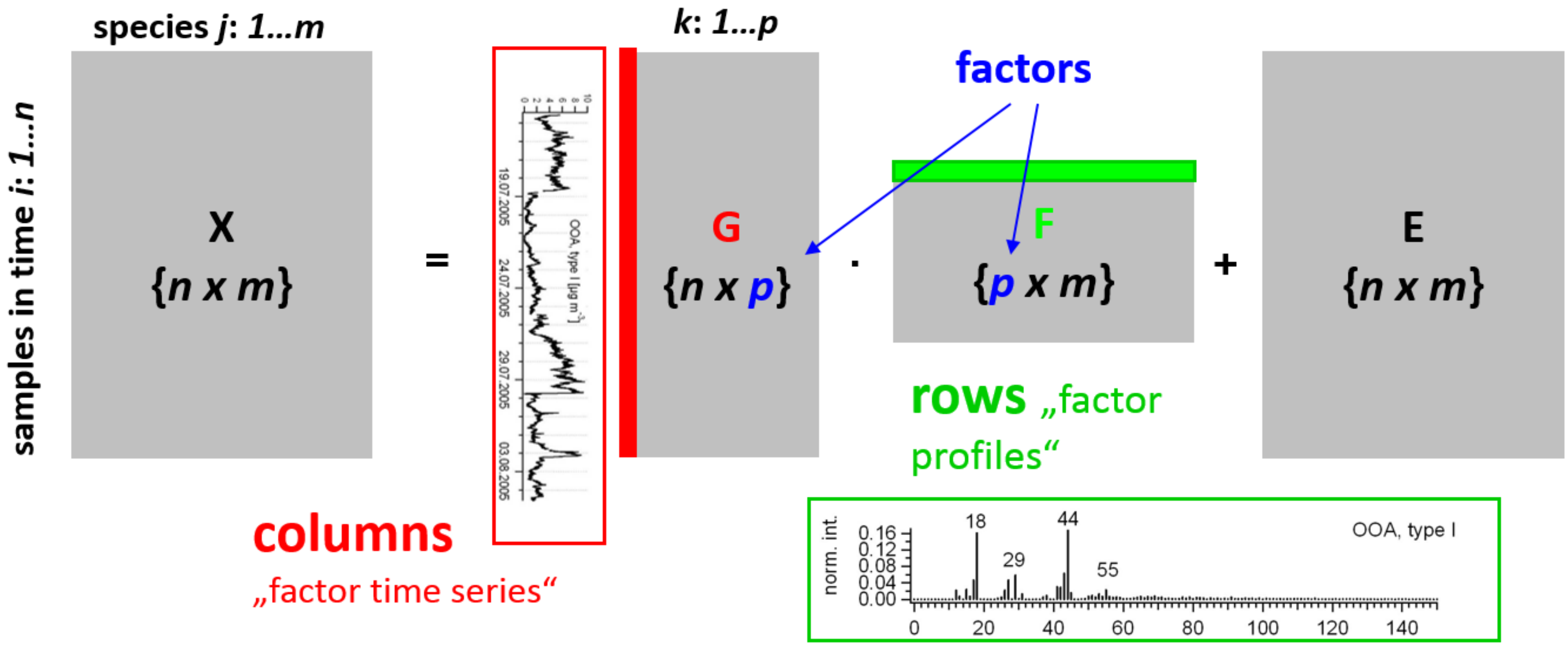
Key words:

PMF, CMB, PMF2, ME-2, Q-space, robust mode, seed runs, local/global minima, rotational ambiguity / uncertainty

Model – Positive Matrix Factorization (PMF)

- **Bilinear factor analytic algorithm**

$$X_{measured} = \hat{X}_{model} + E_{model}$$



Paatero 1994

Model – Positive Matrix Factorization (PMF)

- **Least-squares problem**

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

e_{ij}: difference (measured – model)
σ_{ij}: uncertainty (statistical error)

- Q will be minimized with respect to all model variables
  - ME-2 starts the conjugate gradient algorithm for solving this task

- **Goal**
  - Factor solution must be environmentally reasonable
  - Unstructured residuals over time (ts, diurnals, etc.) and over profile (variables)

Paatero 1999

Model - Q-space

- **Real case**
  - ACSM data with 100 variables for 1000 scans, four factors, unconstrained
  - G{nxp}, F{pxm} ➔ G{1000x4}, F{4x100}, there are 4400 model variables

    ➔ Q(4400 model variables), multidimensional Q-space

- **Simplified case**
  - Simply the real case with two model variables

    ➔ Q(2 model variables), three dimensional Q-space

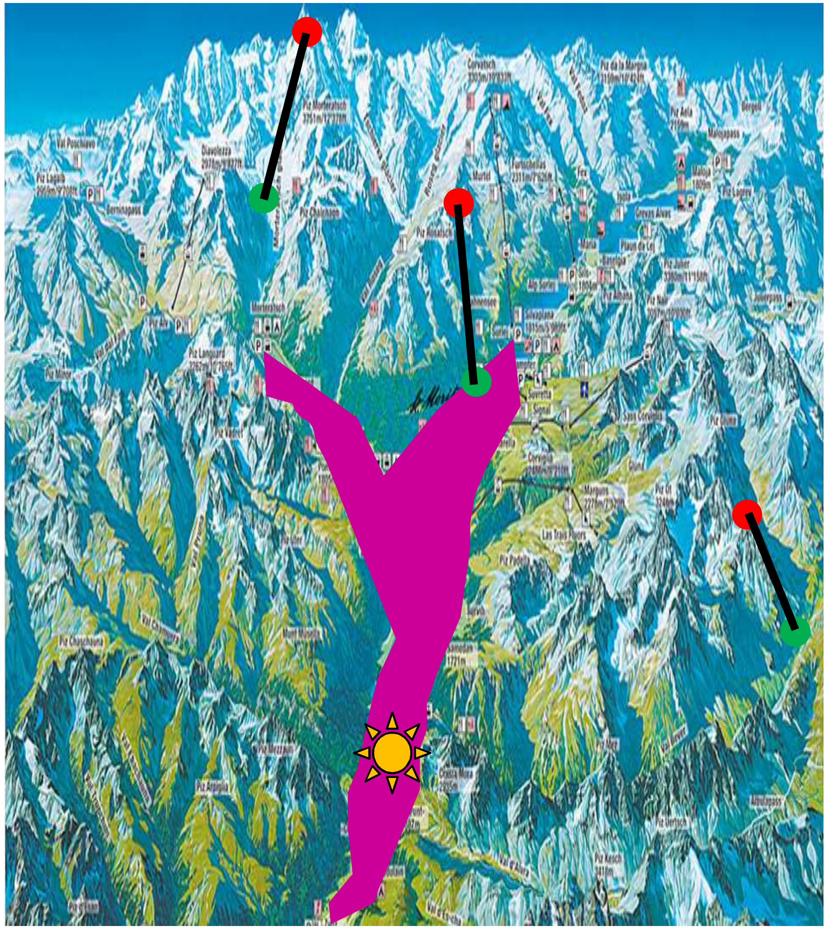
Model - Q-space

- Q(2 model variables) similar to the height h(x,y) in the map
- PMF is performed through the conjugate gradient algorithm minimizing Q based on the starting conditions, following the steepest descent (from red to the green)
- Goal is to find the smallest possible Q-value (global minimum) (violet area) together with the best solution

Search for this minimum based on different starting values (seed run)

- There are many points on the map, for which h(x,y) is equal ➔ rotational ambiguity

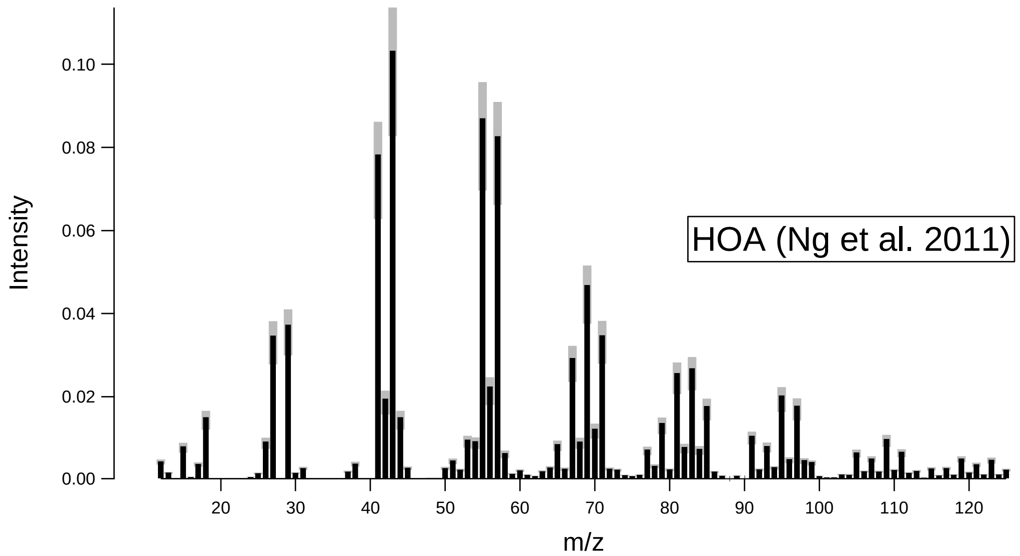
Explore the rotational ambiguity with proper techniques (fpeak, ind. fpeak, a-value, pulling)



Solution space – a-value

- **Assess rotational ambiguity**
- a-value technique
  - Full Q-space can potentially be investigated
  - Advantage: easy to perform and computationally inexpensive
  - Disadvantage: Sensitivity analysis on the constrained model variables

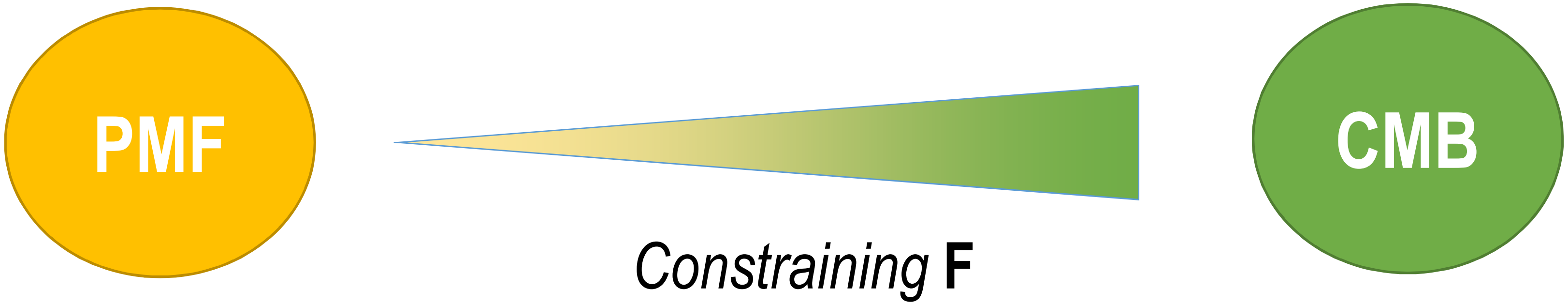
$$f_{p,j,solution} = f_{p,j} \pm a \cdot f_{p,j}$$



Paatero 1999/2008

Model – PMF/CMB/solvers

- **PMF / CMB (chemical mass balance) approach**



- **Solvers**

Solver	Unconstrained	Constrained	Communication
PMF2 / PMF3	X	only to zero	Limited
ME-2	X	X	All quantities easily accessible

Paatero 1999

Model – Positive Matrix Factorization (PMF)

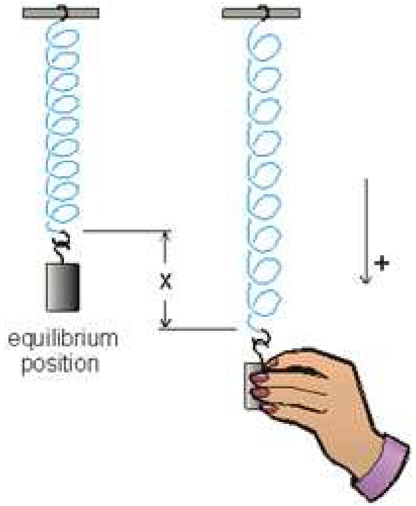
➢ Advantages

– Values in **G** & **F** are non-negative

– Factors represent sources / processes

– PMF algorithm scales with the residual

$$\left(\frac{\partial Q_{ij}}{\partial e_{ij}}\right) \propto e_{ij}$$



$$\left(\frac{\partial E_{pot}}{\partial x}\right) \propto x$$

Paatero 1994

Model – robust mode

➢ PMF run (non-robust mode)

- Computational power is proportional to the residual (in theory ideal)
- Outliers, e.g. transient sources, wrong nb. of factors, electronic recording issues, etc. violate this relation and PMF could spend more time, reducing “wrong” residuals

➢ PMF run (robust mode)

- Allow for this dependency only in a certain range and damp afterwards (robust mode, default value = 4)

$$\text{if}\left|\frac{e_{ij}}{\sigma_{ij}}\right| \leq 4 \;\Rightarrow \left(\frac{\partial Q_{ij}}{\partial e_{ij}}\right) \propto e_{ij} \qquad \text{else}\left|\frac{e_{ij}}{\sigma_{ij}}\right| > 4 \;\Rightarrow \left(\frac{\partial Q_{ij}}{\partial e_{ij}}\right) \propto 4$$

Paatero 1997

Model – Positive Matrix Factorization (PMF)

➢ Disadvantages

- Assess number of factors
- Constant factor profiles (mass spectra)
- Uncertainties are not fully defined, minimal Q-value is not necessarily the best solution

  ➔ Investigate the solution space even for slightly higher Q-values (few %)

- Bilinear factor analytic models suffer from rotational ambiguity

$$\mathbf{X_{model}} = \mathbf{G} \cdot \mathbf{F} = \mathbf{G} \cdot \mathbf{T} \cdot \mathbf{T^{-1}} \cdot \mathbf{F} = \mathbf{G'} \cdot \mathbf{F'}$$

  ➔ Investigate the solution space

Paatero 1994/97

Model – Positive Matrix Factorization (PMF)

➢ **Weight Q by Q_{exp}, the remaining degrees of freedom**

$$Q_{exp} = n \cdot m - p \cdot (n + m) \sim n \cdot m$$

- If all residuals were similar as their σ's, Q / Q_{exp} ~1
- Monitor Q / Q_{exp} values ➔ Too high values may indicate systematic problems of the PMF solution
- Monitor the changes of Q/Q_{exp} over various model runs

V. SoFi workshop
Tutorial – SoFi

Learning goal

-Learn how to prepare the data for a PMF run in SoFi

- Learn how to import and look at various PMF results

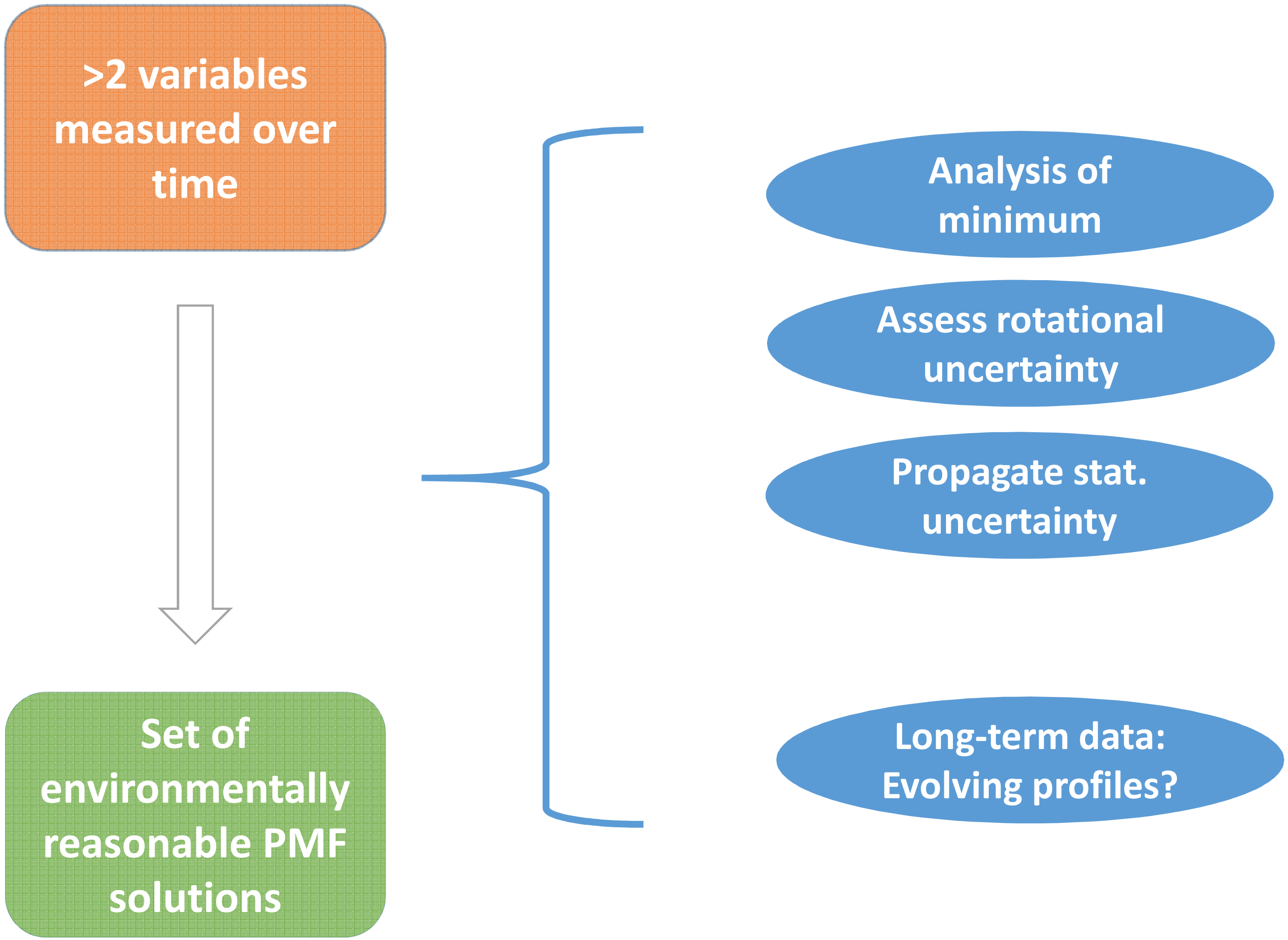
V. SoFi workshop

PMF - advanced

Key words:

ME-2, validation of PMF solution, exploration of solution space (fpeak, a-value, CMB-like approach), propagation of statistical uncertainty, AuRo-SoFi

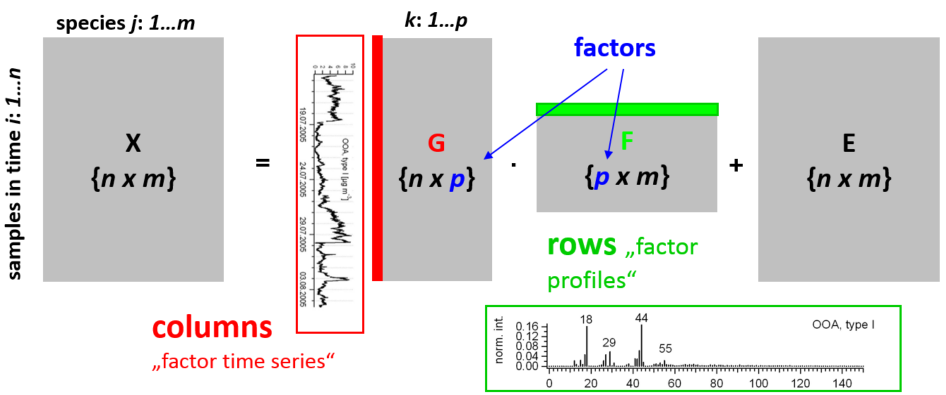
Model – Positive Matrix Factorization (PMF)



Solution space – search for global minimum

➢ **Bilinear factor model (PMF)**

$$X_{measured} = \hat{X}_{model} + E_{model}$$



➢ **Least-squares problem**

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

e_{ij}: difference (measured – model)

σ_{ij}: uncertainty (statistical error)

➢ **Seed runs**

- Initialize PMF run with random values for the unconstrained model variables
- Search for the PMF solution(s) with the smallest possible Q-value (global minimum)

➔compare rotated solutions to this Q-value

Paatero 1994

Solution space – rotational ambiguity

- **PMF solutions suffer from rotational ambiguity**

$$\mathbf{X_{model}} = \mathbf{G} \cdot \mathbf{F} = \mathbf{G} \cdot \mathbf{T} \cdot \mathbf{T}^{-1} \cdot \mathbf{F} = \mathbf{G}' \cdot \mathbf{F}'$$

- **Assess rotational ambiguity**

- Vary the model variables (fpeak, individual fpeak, a-value, CMB-like, pulling) and monitor the change of the PMF solution with various parameters:
  I. Q-value
  II. Residual (global / key variables)
  III. Weighted residual (global / key variables)
  IV. Shape of factor profile(s)
  V. Time series / diurnal correlation with external tracers

Paatero 2008

Solution space – global fpeak

➢ Assess rotational ambiguity

- Global fpeak (ϕ) technique
  - All rotations are performed at the same time
  - Advantage: easy to perform
  - Disadvantage: rotations cannot always be fully predicted, lower estimate of the rotational uncertainty
  - Example:

$$\overline{G} = GT \ \ and \ \ \overline{F} = T^{-1}F \qquad T_{\text{fpeak, p=3}} = \begin{bmatrix} 1 & \phi & \phi \\ \phi & 1 & \phi \\ \phi & \phi & 1 \end{bmatrix}$$

Paatero 2008

Solution space – individual fpeak

➢ Assess rotational ambiguity

- Individual fpeak (ϕ) technique
  - All rotations are performed at the same time
  - Advantage: easy to perform
  - Disadvantage: rotations cannot always be fully predicted, lower estimate of the rotational uncertainty
  - Example:

$$\overline{G} = GT \ \ and \ \ \overline{F} = T^{-1}F \qquad T_{p=3} = \begin{bmatrix} 1 & 0 & \phi \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

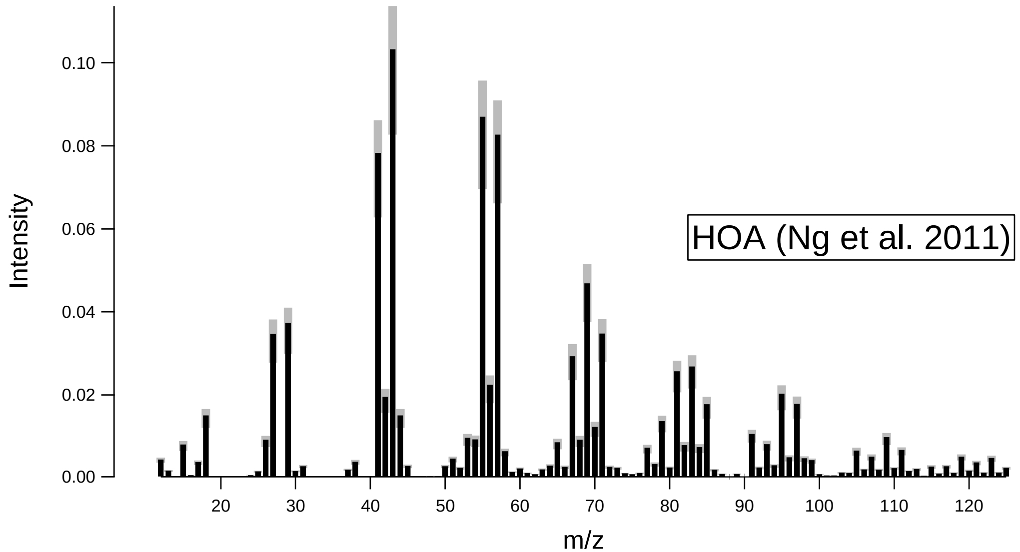
Paatero 2008

Solution space – a-value

➢ **Assess rotational ambiguity**

- a-value technique
  - Full Q-space can potentially be investigated
  - Advantage: easy to perform and computationally inexpensive
  - Disadvantage: Sensitivity analysis on the constrained model variables

$$f_{p,j,solution} = f_{p,j} \pm a \cdot f_{p,j}$$



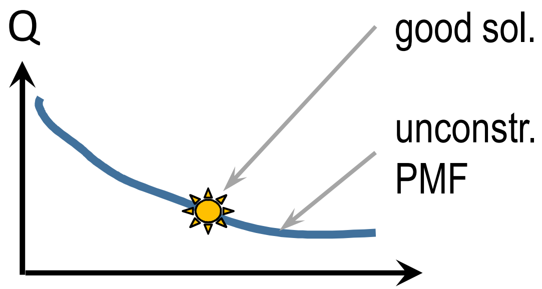
Paatero 1999/2008

Solution space – a-value

➢ Assess rotational ambiguity

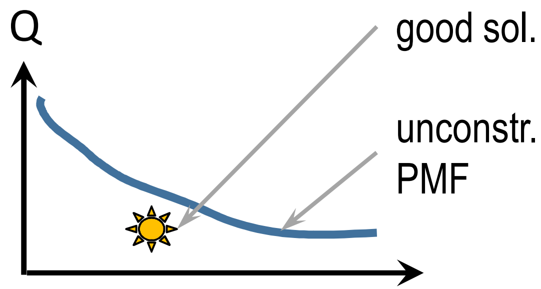
- a-value technique
  - Full Q-space can potentially be investigated
  - Advantage: easy to perform and computationally inexpensive
  - Disadvantage: Sensitivity analysis on the constrained model variables

Good case



Sensitivity analysis performed on the constrained anchor meets/finds the good solution

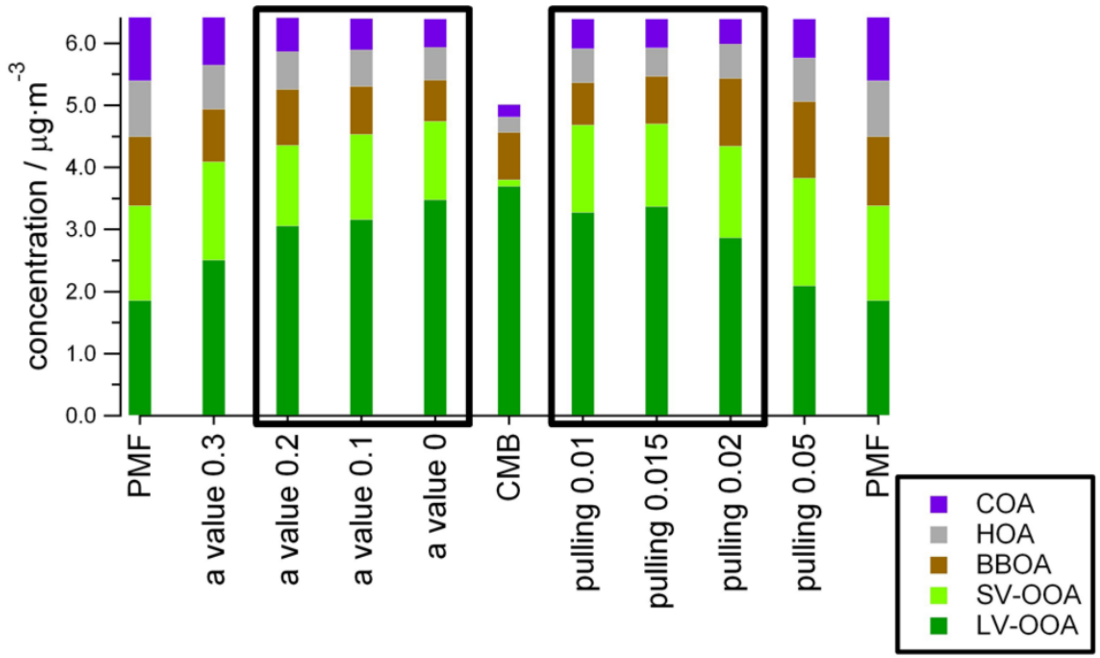
Bad case



Sensitivity analysis performed on the constrained anchor does not find the good solution
➔ change factor profile (**AMS Spectral Database**)
(http://cires.colorado.edu/jimenez-group/AMSsd/)

Solution space – a-value

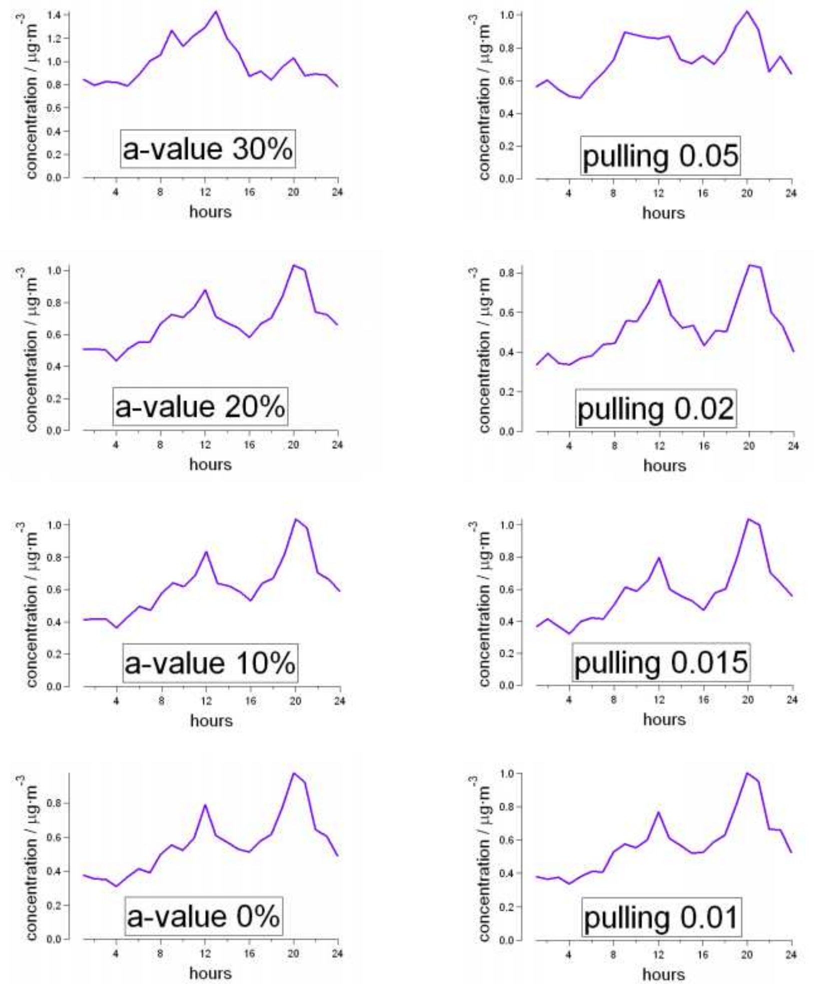
- **ACSM Zurich winter 2011**
  - Employed anchors (HOA, COA, BBOA) meet reasonable solutions for a-value range 0 – 0.2

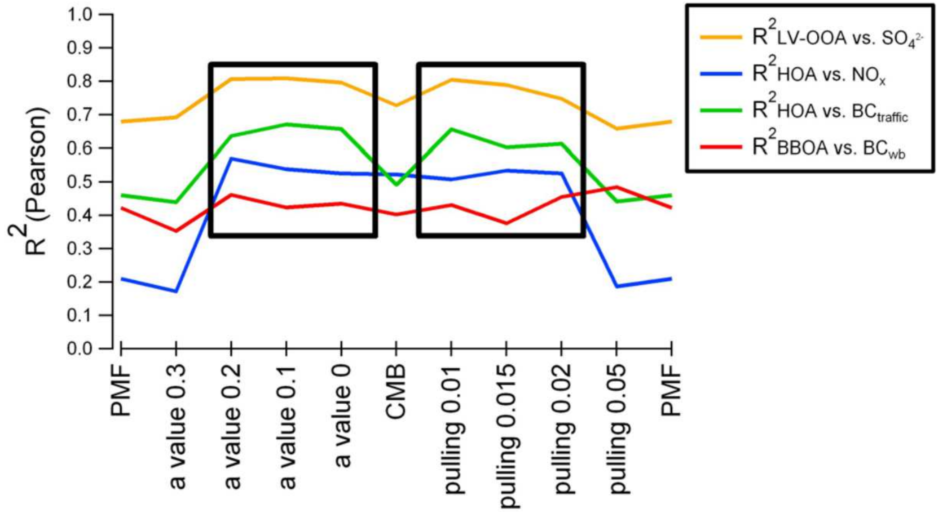


Canonaco et al. 2013

Solution space – a-value

- **ACSM Zurich winter 2011**

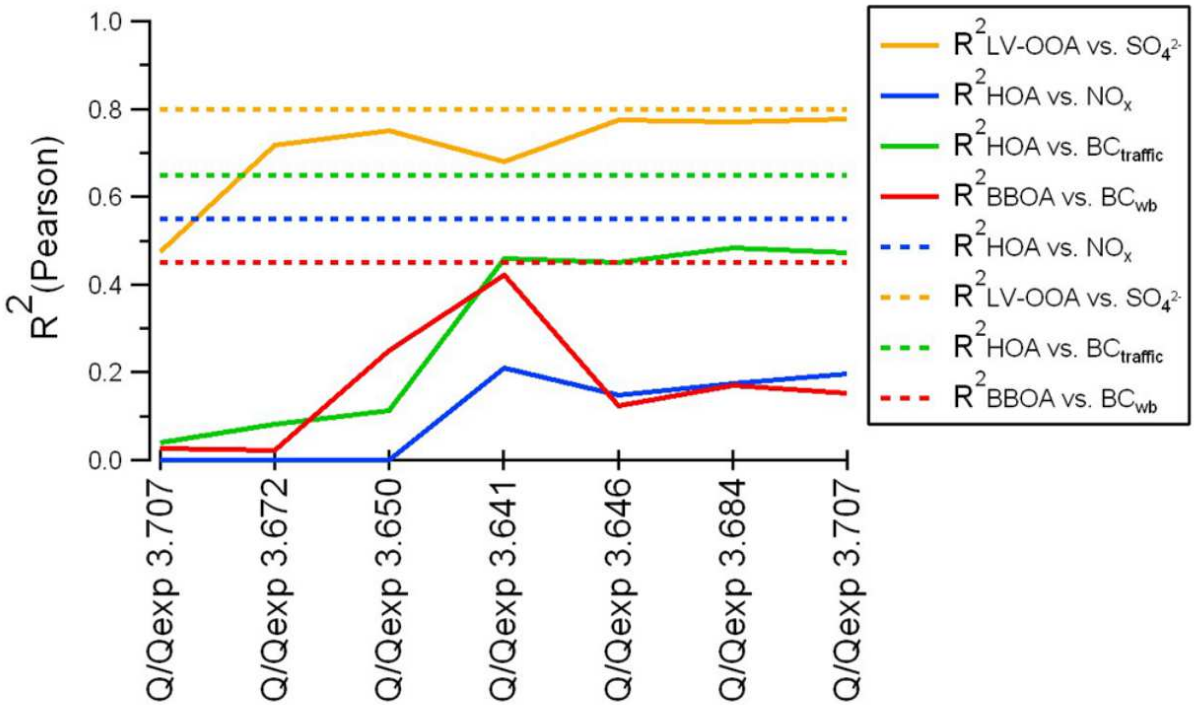




Canonaco et al. 2013

Solution space – comparison to fpeak

- **ACSM Zurich winter 2011**
  - Pure fpeak analysis (solid lines) do not reproduce the reasonable PMF solutions (dashed lines)



Canonaco et al. 2013

Solution space – a-value space investigation

➢ **Issue for a-value space for two and more constrained vectors**

- Analyzing systematically all solutions becomes difficult
- Possible alternative appraoch

  ➔ reorder all PMF solutions based on a list of possible criteria

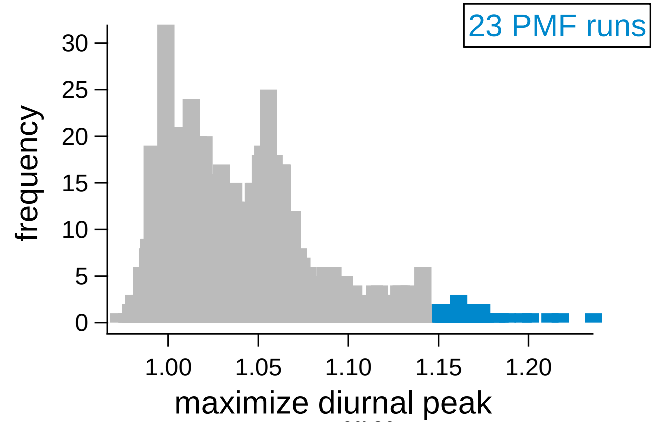
  ➔ dynamic change of the criteria/weight and inspection of the PMF solutions

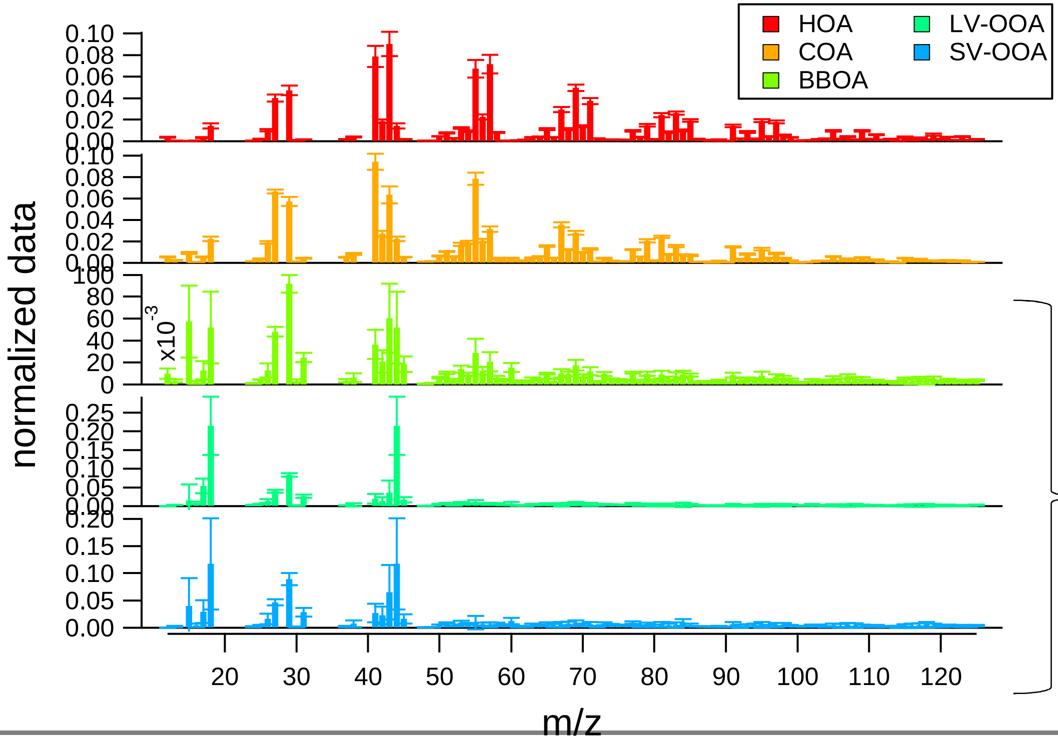
- Package (criteria-based approach) ready to the shared for testing (free license for one year)
- More details presented later this morning (Yuliya Sosedova)

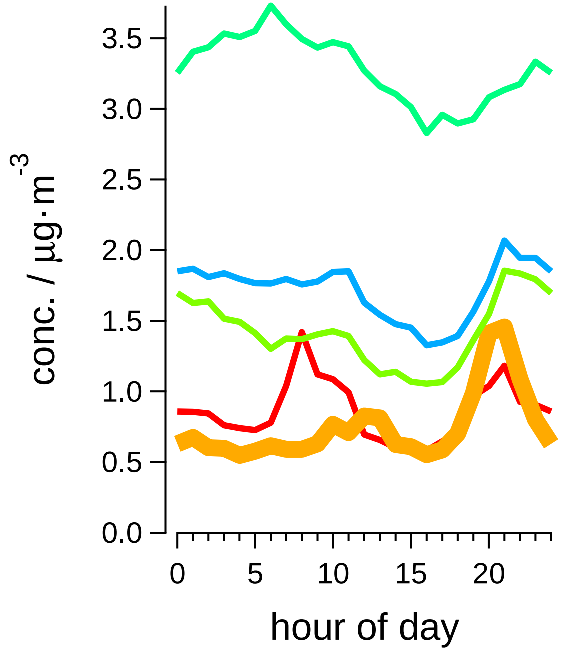
Canonaco et al. 2013

Solution space – a-value space investigation

factors	criteria
constrained HOA	ts-correlation NO_x
constrained COA	**maximize diurnal peak at noon**
constrained BBOA	ts-correlation to BC_{wb}
LV-OOA	ts-correlation to sulfate
SV-OOA	ts-correlation to nitrate







Canonaco et al., in prep.

Solution space – statistical error prop.

- **Propagate the statistical uncertainty to the PMF result**
  - Monte Carlo method (noise insertion)

    vary the PMF input within the statistical error and call PMF
  - Bootstrap method (resampling strategy)

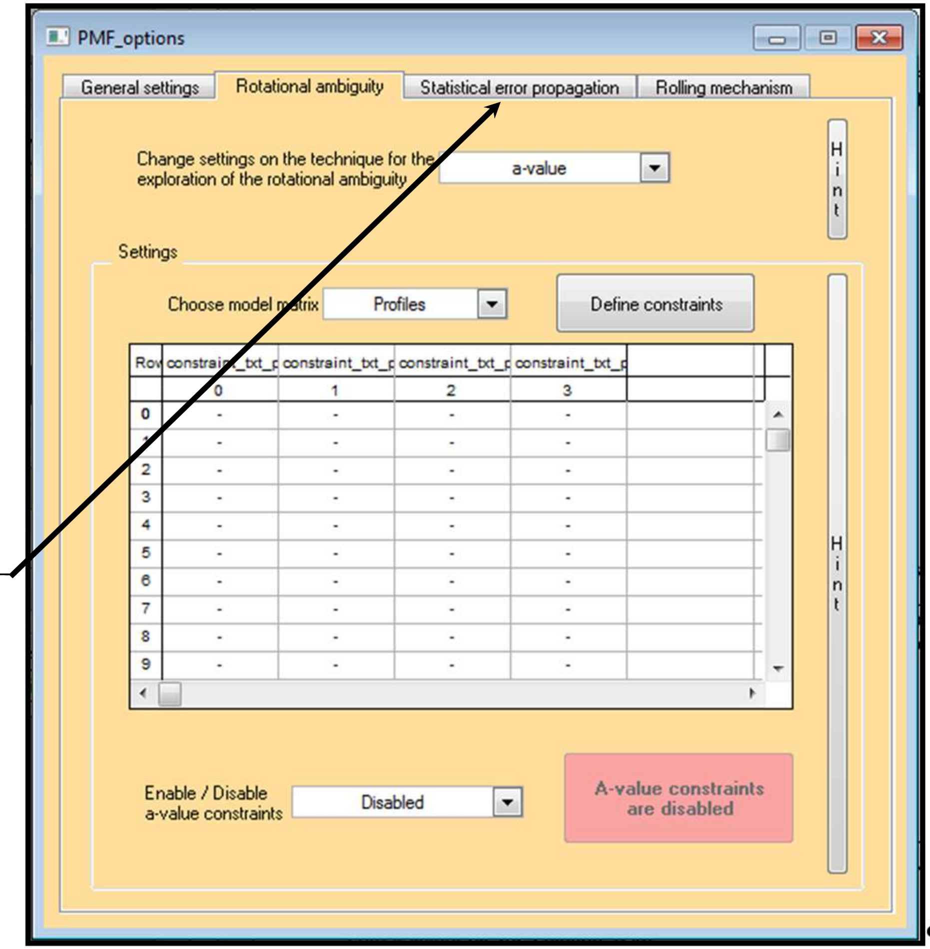
    resample data with identical underlying sources and call PMF

Solution space – statistical error prop.

➢ **Propagate the statistical uncertainty to the PMF result**

Approach should be performed in addition to the exploration of the rotational ambiguity)

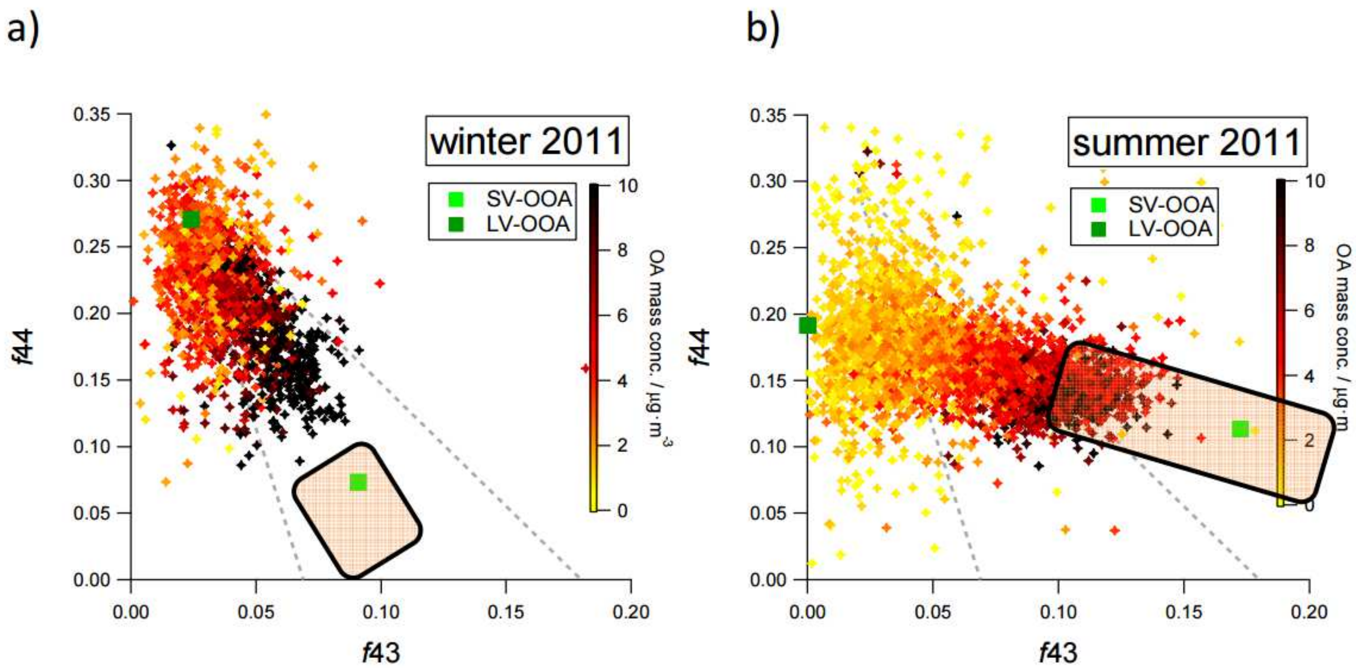
I. Select approach
II. Enable
III. Run PMF (e.g. 10-1000x more runs)
IV. Analyze runs already containing the statistical uncertainty



Solution space – AuRo-SoFi

- **ACSM Zurich: winter and summer data 2011**
  - SOA f44/f43 vary over the seasons
  - Running PMF over the entire year would average this out

➔ Run PMF season/month-wise (manually) / apply AuRo-SoFi (automatic)



Canonaco et al. 2015

Solution space – AuRo-SoFi

➢ AuRo-SoFi algorithm

- Run PMF using a small frame, e.g. two weeks/one month of data (Assumption: source is constant over this period)
- Optimize PMF solution based on criteria defined in advance based on manual pretests
  ➔Automatic part
- Shift PMF frame forward and rerun PMF
- Repeats for small shifts (daily shift compared to length of PMF frame) is facsimile of the bootstrap technique and hence partially propagates the stastistical uncertainty
  ➔ Rolling part

➔ **AuRo – SoFi algorithm**

- More details presented on Wednesday afternoon (Yuliya Sosedova)

Canonaco et al. in prep., Sosedova et al., in prep.