PAUL SCHERRER INSTITUT

PSI

Paul Scherrer Institute

SoFi workshop

Prague, 15. – 17.2. 2018

Presented by Amelie Bertrand, Giulia Stefenelli and Anna Tobler

General – Program and logistics

Thursday	
Time	*Activity*
9:00 – 10:30	• General information • Presentations of the participants
	+++ COFFEE BREAK +++
10:50 – 12:30	• Presentations of the participants
	+++ LUNCH +++
13:30 – 15:00	• Software installation if not done so ealier • Theory input on PMF, ME-2, Q-space, robust mode, rotational tools (a-value)
	+++ COFFEE BREAK +++
15:30 – 17:00	• Guide through SoFi, main features (import raw data, treat data for PMF run, call ME-2)

- Download the software and sample data from
  https://www.psi.ch/acsm-stations/sofi-training-school-2018

General – Program and logistics

Friday	
Time	*Activity*
9:00 – 10:30	• Explore PMF results in SoFi
	+++ COFFEE BREAK +++
10:45 – 12:30	• Explore PMF results in SoFi
	+++ LUNCH +++
13:30 – 15:00	• Individual work: participants work on their own data set (support provided)
	+++ COFFEE BREAK +++
15:30 – 17:00	• Individual work: participants work on their own data set (support provided) • Group work: users treating similar data have the chance to share gained experience

General – Program and logistics

Saturday	
Time	*Activity*
9:00 – 10:30	• Outlook SoFi-Pro • Presentations from PSI • Conclusion
	+++ COFFEE BREAK +++
10:45 – 12:30	
	+++ LUNCH +++
13:30 – 15:00	• Open discussions on results of the participants

General – Presentation of the participants

Time	Participants	Institute
	Vincent Crenn	Addair
	Leah Williams	Aerodyne Research, Inc.
	Lütfi Peker	Department for Environmental Science Aarhus University
	Ettore Petralia	ENEA Bologna
	Zoltan Nemeth	Eötvös University
	Aku Helin	FMI ACR
	Tanguy Amodeo	INERIS
	Cristina Antonia Marin	INOE
	Otakar Makes, Petra Pokorna, Hichem Bouzidi	Institute of Chemical Process Fundamentals of the CAS
	Anka Cvetkovic	Institute of Public Health of Belgrade
	Maria Fernández-Amado	Institute Universitary of Environment, UDC
	Antoine Farah	LAMP
	Benjamin Chazeau	LCE
	Athina-Cerise Kalogridis	NCSR Demokritos
	Elena Hristova	NIMH
	Ana Cvitesic Kusan	Ruder Boskovic Inst.
	Iasonas Stavroulas	The Cyprus Inst.
	Magdalena Kistler	TU Wien
	Joana Lage, Nuno Henrique Varela Canha	Universidade de Lisboa
	Aikaterini Bougiatioti	University of Crete
	Marija Zivkovic	Vinča Institute of Nuclear Sciences

PAUL SCHERRER INSTITUT

PSI

Prague, 15. - 17.2.2017

PMF – general

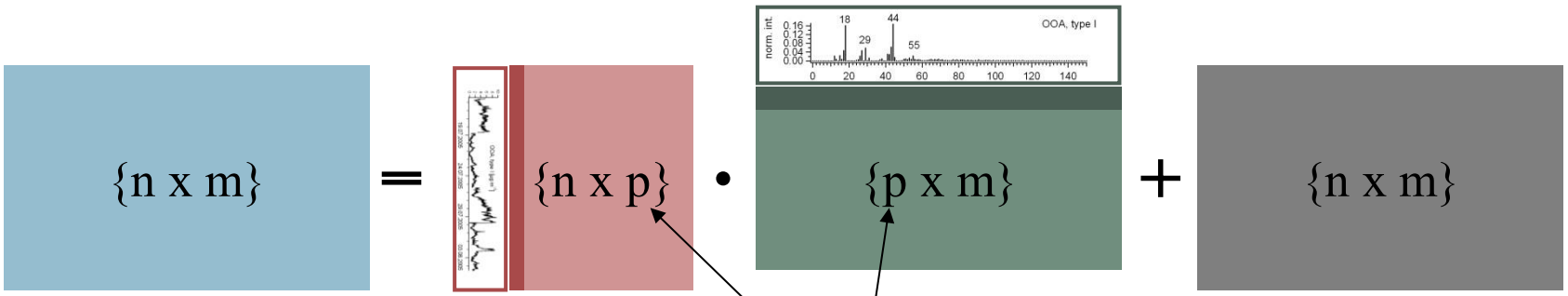
Keywords:
PMF, CMB, PMF2, ME-2, Q-space, robust mode, seed runs, local/global minima, rotational ambiguity / uncertainty

Model - Positive Matrix Factorization (PMF)

Bilinear factor analytic algorithm

$$X_{measured} = G \cdot F + E$$

mass spectral matrix | factor contribution | factor profile | residual matrix



Values in G & F are non-negative
Factors represent sources / processes

factors

Goal
Factor solution must be environmentally reasonable

Paatero, P., Tapper, U., 1994. Environmetrics 5, 111-126.

Model – Positive Matrix Factorization (PMF)

Least-square problem

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

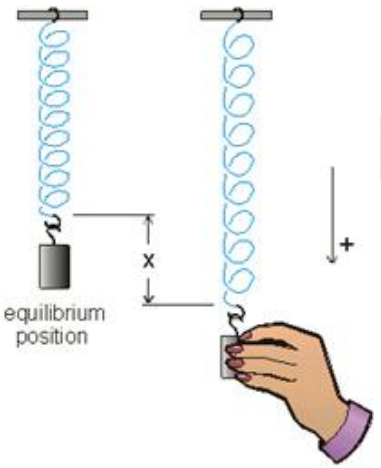
e_{ij}: difference (measured – model)

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 (multilinear regression)

- Quantity minimized in ME-2 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, P., 1997. Chemometr Intell Lab 37, 23-35.

PMF – robust mode

PMF run (non-robust mode)

- Minimized quantity 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" entries in **Q**

PMF run (robust mode)

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

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

PMF – Q/Q_{exp}

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 might indicate systematic problems of the PMF result
- Monitor the changes of Q/Q_{exp} over various model runs

PMF - problems

Disadvantages

- Assess number of factors
- Constant factor profiles (mass spectra)
- Uncertainties are not fully defined, minimal Q-value is not necessarily the best solution
- Bilinear factor analytic models suffer from rotational ambiguity

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

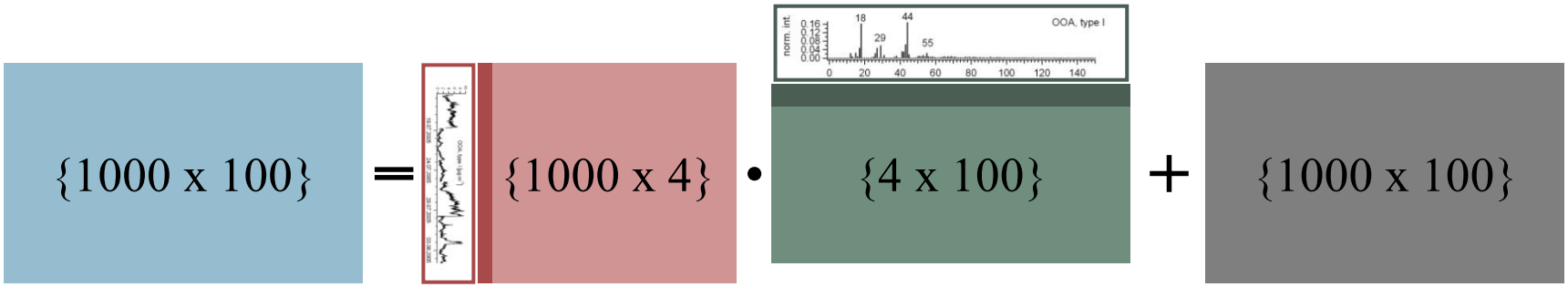
→ explore the PMF results

– vary number of factors

– vary the entries in G and F randomly (seed), controlled (e.g. a value, if good reasons) to find the global minimum/minima

PMF – Q-space

- Real case
  - ACSM data with 100 variables for 1000 scans, four factors, unconstrained



→ 4400 model variables

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

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

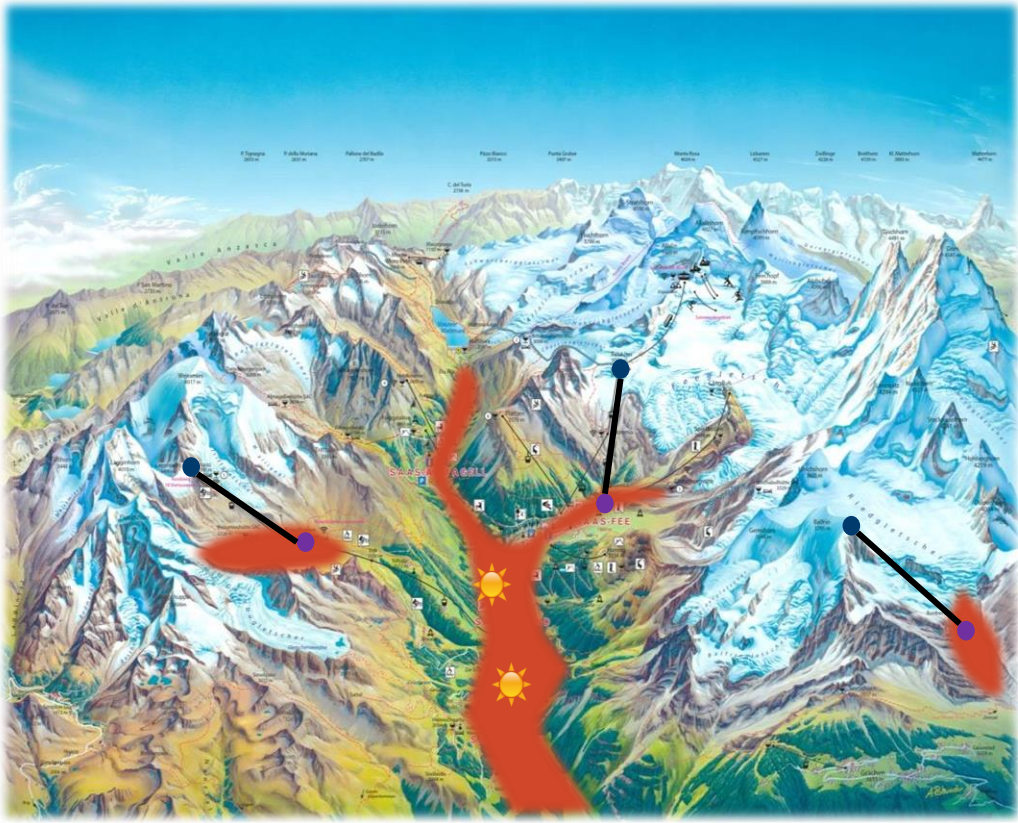
→ Q(2 model variables), three dimensional Q-space

PMF – rotational ambiguity

- 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 blue to the purple)

- Goal is to find the smallest possible Q-value (global minimum) (red area) together with interpretable PMF runs (☀)

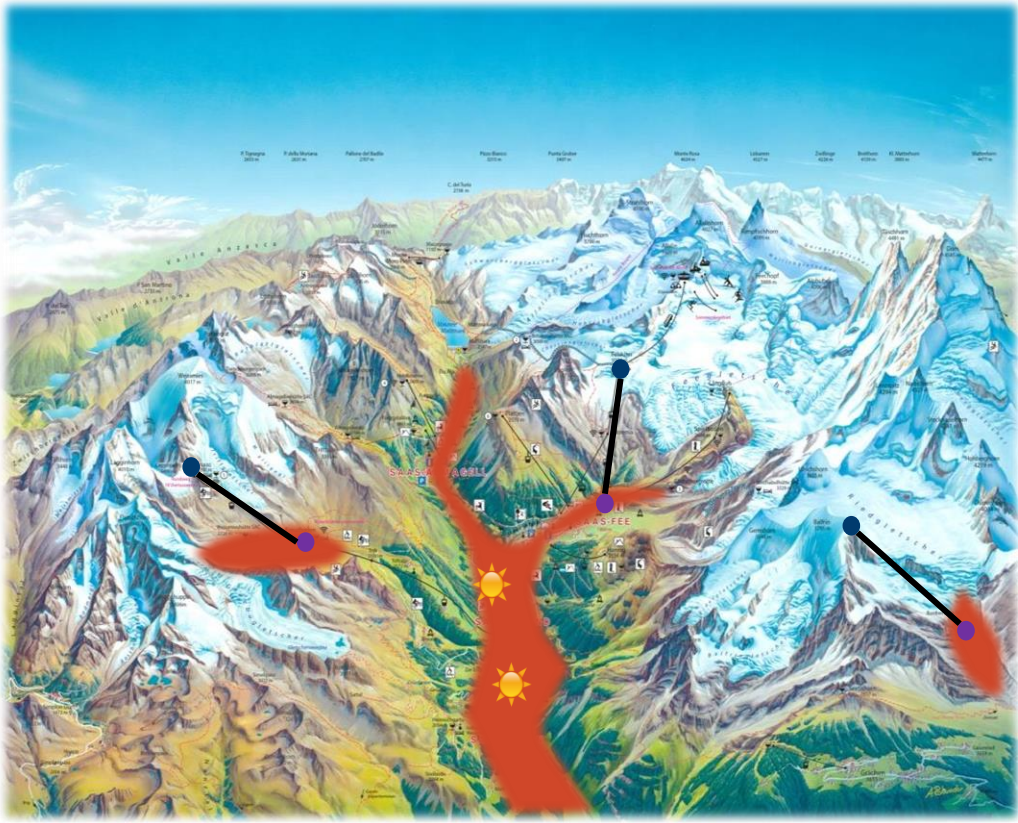


PMF – rotational ambiguity

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

- the task is extremely difficult using only complete random entries in **G** and **F** → could take lots of PMF runs

- Explore the rotational ambiguity with proper techniques (global fpeak, individual fpeak, a-value, CMB-like)

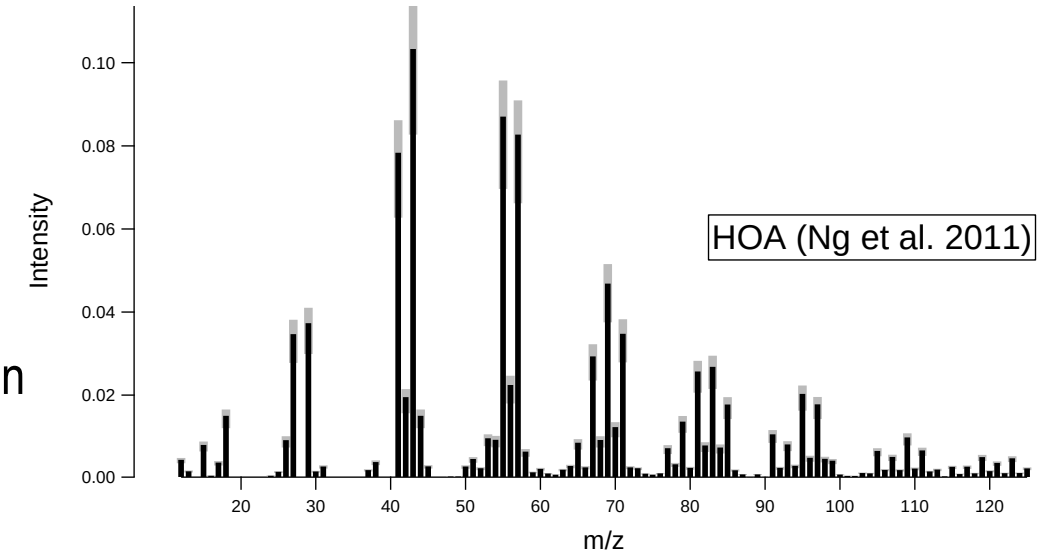


PMF – a-value technique

- full Q-space can potentially be investigated
- advantage: easy to perform and computationally inexpensive
- if **a priori information** available, wiser to confine solution space by constraining entries in **G** and / or **F**
- disadvantage: sensitivity analysis on the constrained model variables

PMF – a-value technique

- full Q-space can potentially be investigated
- advantage: easy to perform and computationally inexpensive
- if **a priori information** available, wiser to confine solution space by constraining entries in **G** and / or **F**
- disadvantage: sensitivity analysis on the constrained model variables



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

PMF – a-value technique

- a value approach allows to move in specific directions that would be forbidden in a pure unconstrained / seed run (rectangle)

- the a value sensitivity analysis performed on the constrained anchor meets/finds the good solution (solid black line)

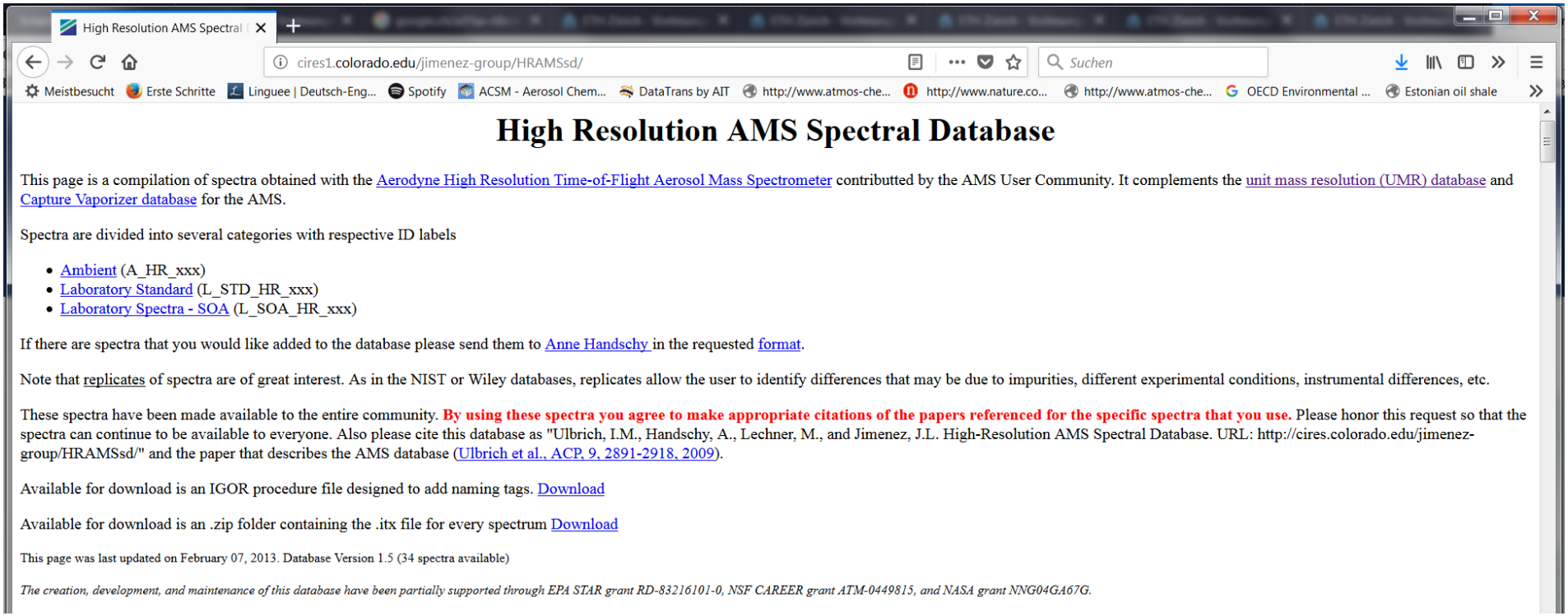
PMF – a-value technique

- sensitivity analysis performed on the constrained anchor does not find the good solution

- change position of initial value (dashed line), i.e., change factor profile

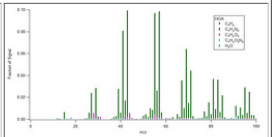
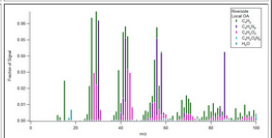
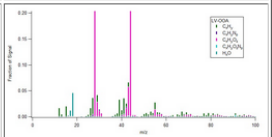
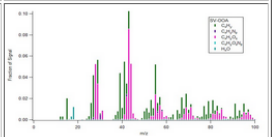
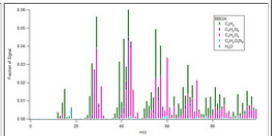
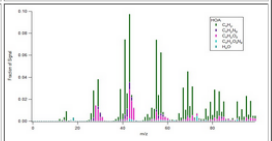
PMF – a priori information

- ~ a decade of know-how on AMS fingerprints of various POA & SOA, exploit this information in PMF
- HR and UMR profiles available from Jimenez' website



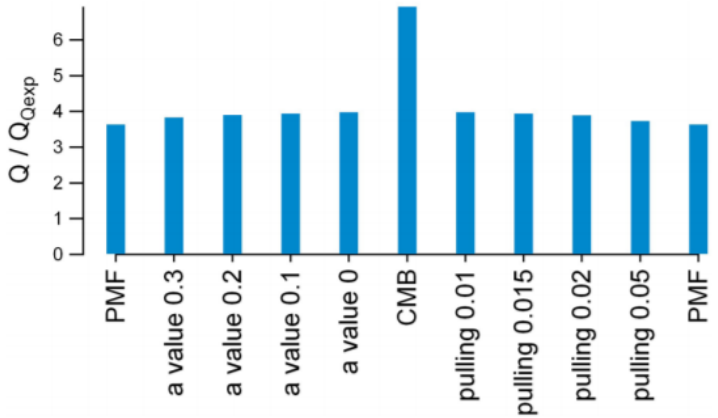
PMF – a priori information

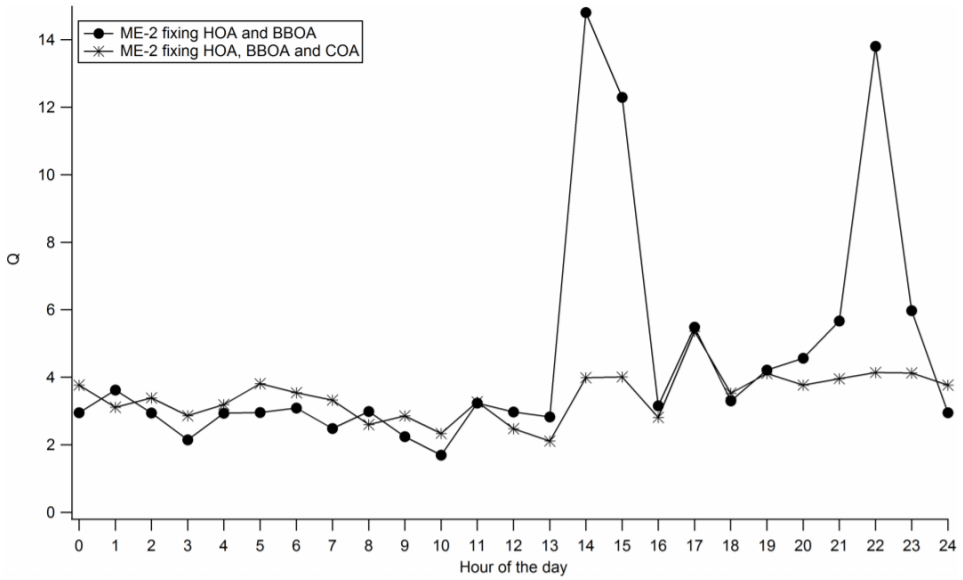
Ambient

Spectra ID	Source	Group	AMS Instrument	EI Energy	Vaporizer Temp (°C)	Citation	Fig #	Comments	Data
A_HR_001	HOA	K.S. Docherty Jimenez – University of Colorado	HR-ToF (V mode)	70 eV	600	Docherty, K.S., et al., **ACP** 2011, *11*, 12387-12420.	Fig 7, pg. 12403	SOAR-1 Campaign, Riverside, CA, July 2005	 A_HR_001_HOA.itx
A_HR_002	LOA	K.S. Docherty Jimenez – University of Colorado	HR-ToF (V mode)	70 eV	600	Docherty, K.S., et al., **ACP** 2011, *11*, 12387-12420.	Fig 7, pg. 12403	SOAR-1 Campaign, Riverside, CA, July 2005	 A_HR_002_LOA.itx
A_HR_003	LV_OOA	K.S. Docherty Jimenez – University of Colorado	HR-ToF (V mode)	70 eV	600	Docherty, K.S., et al., **ACP** 2011, *11*, 12387-12420.	Fig 7, pg. 12403	SOAR-1 Campaign, Riverside, CA, July 2005	 A_HR_003_LV_OOA.itx
A_HR_004	SV_OOA	K.S. Docherty Jimenez – University of Colorado	HR-ToF (V mode)	70 eV	600	Docherty, K.S., et al., **ACP** 2011, *11*, 12387-12420.	Fig 7, pg. 12403	SOAR-1 Campaign, Riverside, CA, July 2005	 A_HR_004_SV_OOA.itx
A_HR_005	BBOA	A.C. Aiken Jimenez – University of Colorado	HR-ToF (V mode)	70 eV	600	Aiken, A.C., et al., **ACP** 2009, *9*, 6633-6653.	Fig 3, pg. 6641	MILAGRO Campaign T0, March 2006	 A_HR_005_BBOA.itx HR 4 PMF factors (.xlsx file)
A_HR_006	HOA	A.C. Aiken Jimenez – University of Colorado	HR-ToF (V mode)	70 eV	600	Aiken, A.C., et al., **ACP** 2009, *9*, 6633-6653.	Fig 3, pg. 6641	MILAGRO Campaign T0, March 2006	 A_HR_006_HOA.itx

PMF results – goals

- **PMF results must be**
  - mathematically acceptable (similar Q, unstructured scaled residuals over **time** (ts, diurnal cycle, weekly cycle, etc.) and over **profile** (variables))



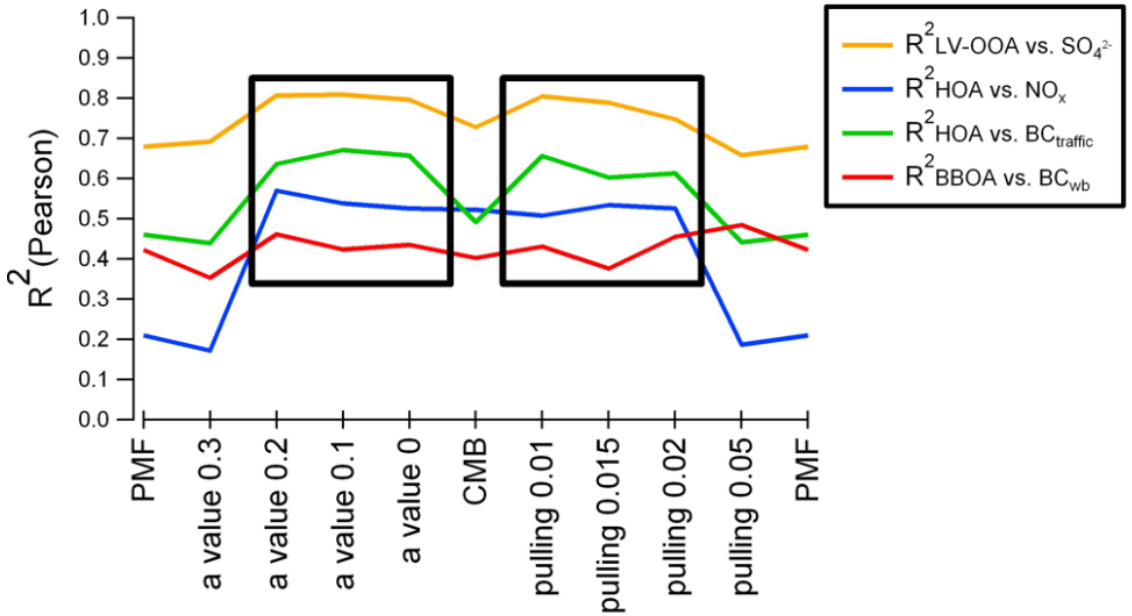


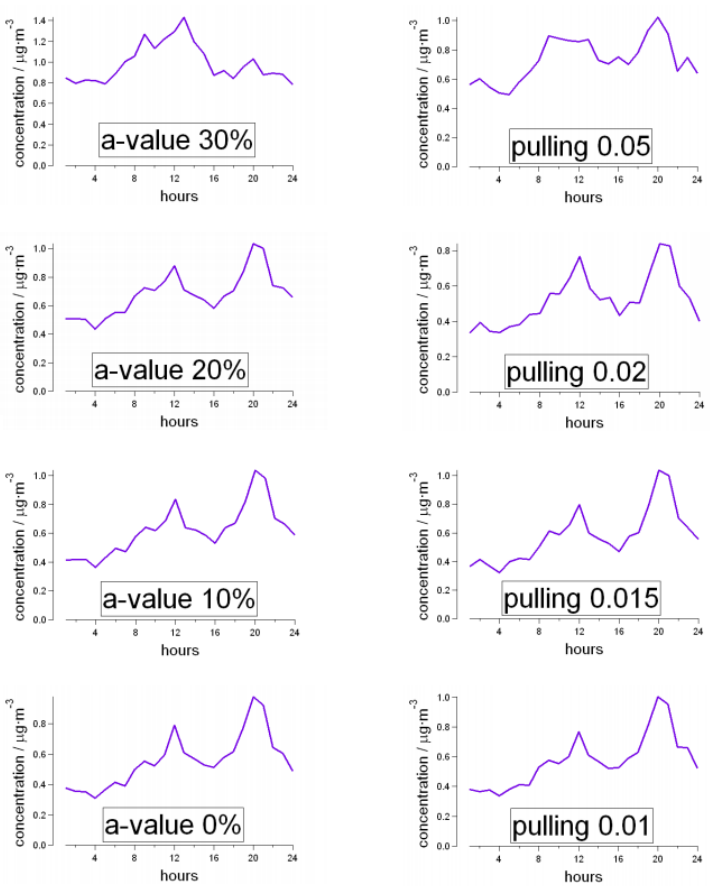
Canonaco et al. 2013

Crippa et al. 2014

PMF results – goals

- **PMF results must be**
  - factor solution must be environmentally reasonable

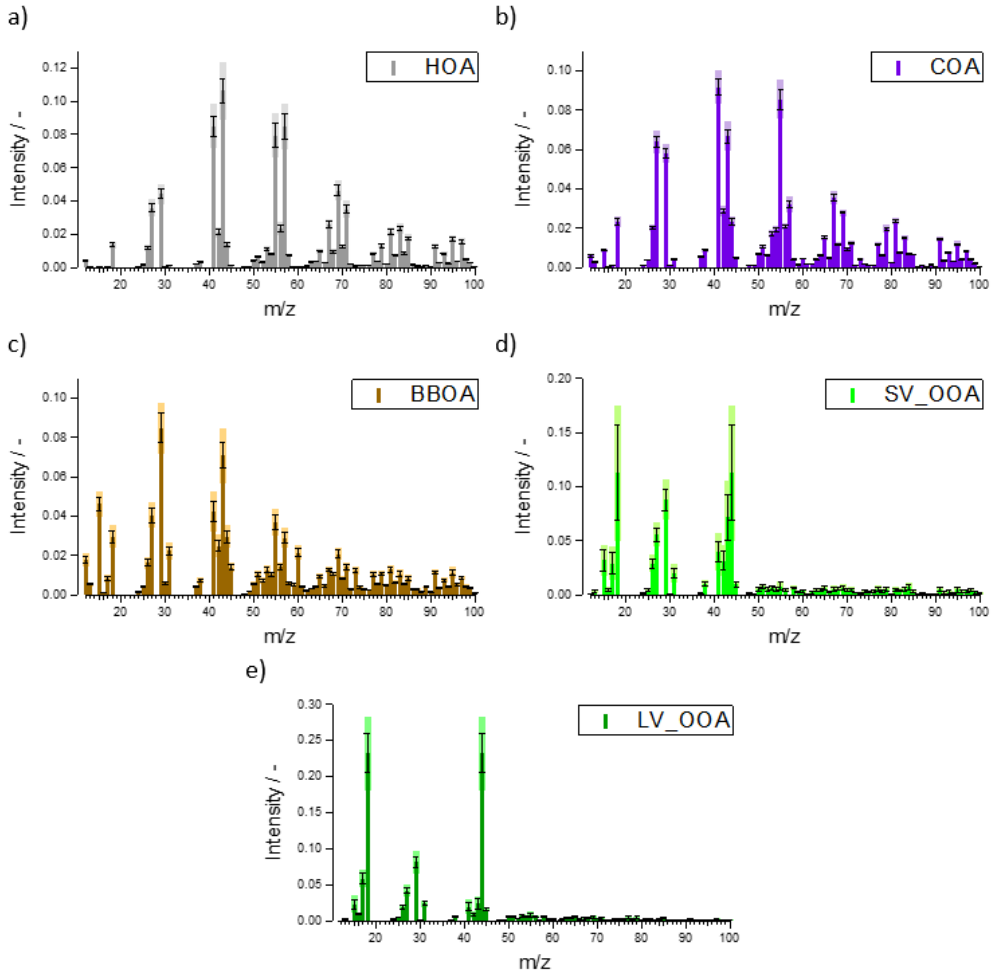




Canonaco et al. 2013

PMF – uncertainties of PMF solutions

- **PMF solution**



Canonaco et al. in prep

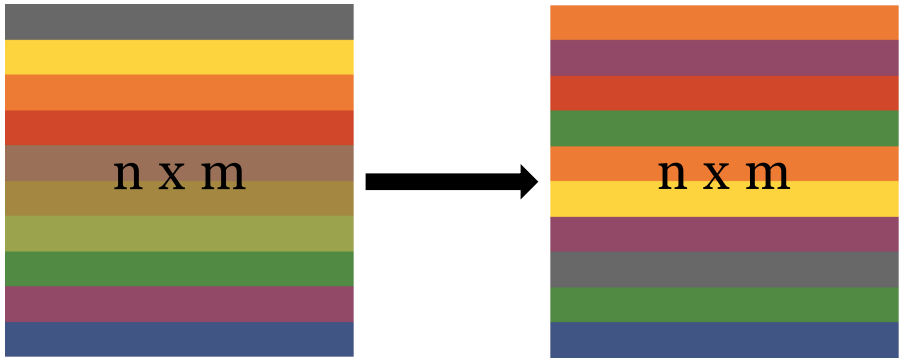
PMF – uncertainties of PMF solutions

- **PMF solution**
  - PMF solution contains all PMF results that are environmentally reasonable (!not only one PMF run!)
  - rotational uncertainty (amount of rotational ambiguity) is assessed using e.g. the a value technique
  - statistical uncertainty, for ambient ACSM data mainly the daily variation of the sources, is assessed using the resampling strategy «bootstrap»
- **More details on Saturday**

PMF - bootstrap

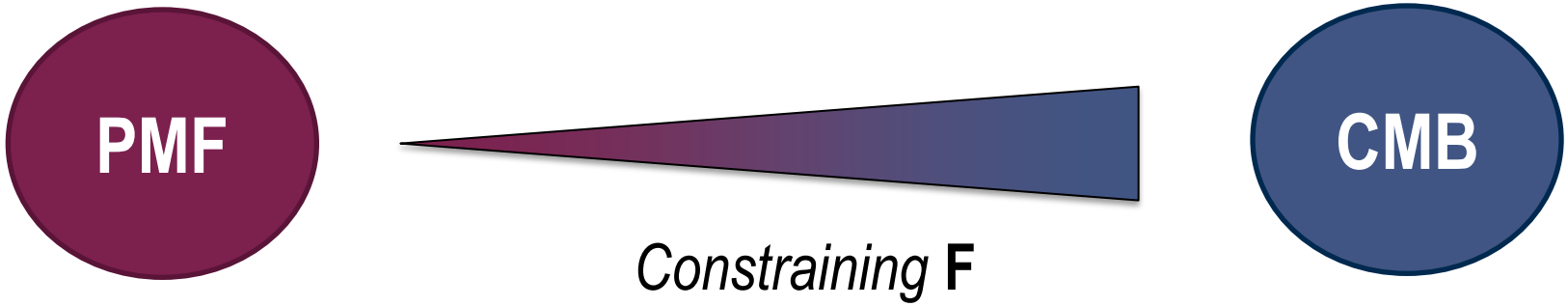
- Uncertainty estimation / variability in PMF solution
  - Randomly selecting rows or blocks of consecutive samples
  - Create new data set with dimensions of the original data set
  - PMF runs on resampled data set

- Does not account for rotational ambiguity



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, P., 1999. J Comput Graph Stat 8, 854-888.

PAUL SCHERRER INSTITUT

PSI

Prague, 15. - 17.2.2017

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

General information

- Current policy using SoFi
  - collaboration for at least two peer-reviewed manuscripts per scientific group (F. Canonaco, A. Prevot and PSI people supporting your analysis during the workshop)
  - cite the SoFi paper in AMT (Canonaco et al. 2013)

- standard SoFi software
  - freeware (only collaboration)
  - relevant information is posted to the google group “SoFi_ME2” (to join contact Francesco Canonaco, francesco.canonaco@psi.ch)
  - as a member you will receive the password for extracting the latest SoFi software from the PSI homepage:
  - https://www.psi.ch/acsm-stations/me-2

- SoFi Pro
  - license-based
  - first official release this summer
  - contact Andre Prevot / Francesco Canonaco

General information

IGOR-based

- currently working with IGOR 6.36
- will be made compatible with IGOR 7 starting from this spring/summer

Concept in SoFi

- exploration base case using a priori information, constrain what is known to find out what is unknown (iterative way)
- assess uncertainties rotational using random a values and statistical using bootstrap

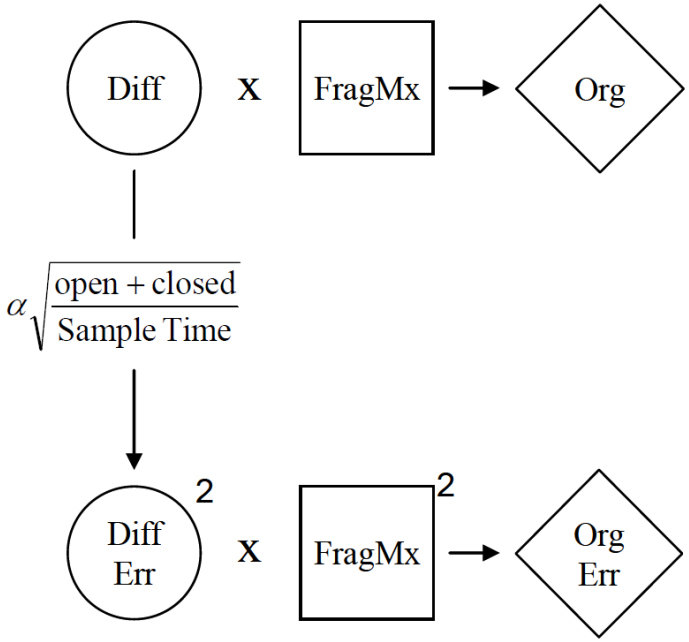
Required information

SoFi will need the following input

- data matrix
- error matrix
- numeric wave that has index values for your species (1, 2, 3, ...)
  - text wave that has the names of your species (short names are better) but can be created by SoFi
- a time wave

PMF – Error matrix

- Standard AMS error calculation



Allan et al., 2003

- Direct calculation in the softwares for AMS and ACSM

PMF – Error matrix

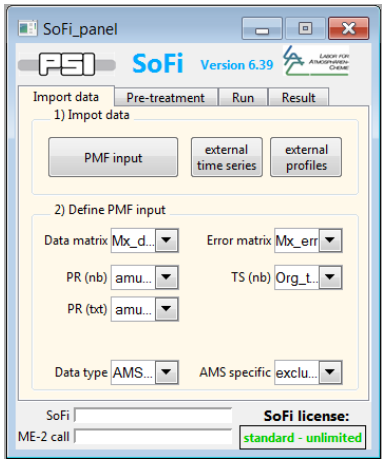
- **PTR Data Set**

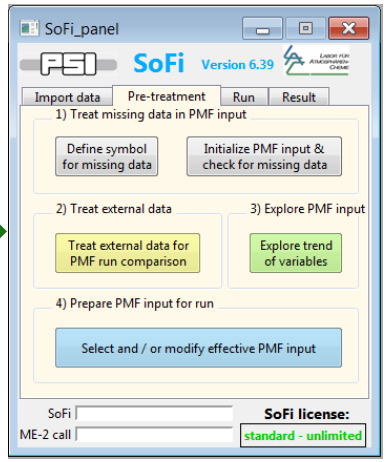
$$Error = \frac{1}{I_{H3O+} * sensitivity} * \sqrt{\frac{I_{meas.}}{dwelltime_{meas.}} + \frac{I_{background.}}{dwelltime_{background.}}}$$ Crippa et al. (ACP, 2013)

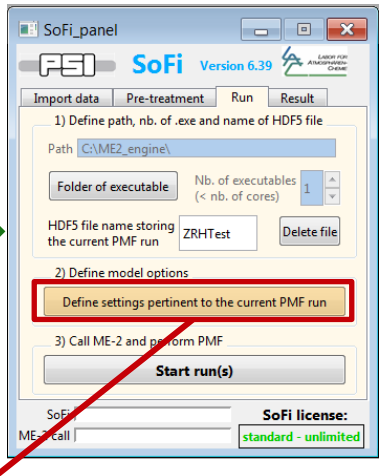
- **GC/MS Data Set (2 methods)**
  - **Polissar et al. (JGR, 1998)**
    - If C < DL replace C with LD/2 and $Error = \frac{5}{6} LD$
    - If C > DL $Error = \frac{1}{3} LD$
  - **Gianini et al. (Atmo Envi., 2012)**
    - $Error = \sqrt{DL^2 + (CV * C)^2 + (a * C)^2}$

      where $a = 0.03$ and CV is the Standard Deviation average of replicates

SoFi 6.39 – prepare PMF input







Define input data
Add external data

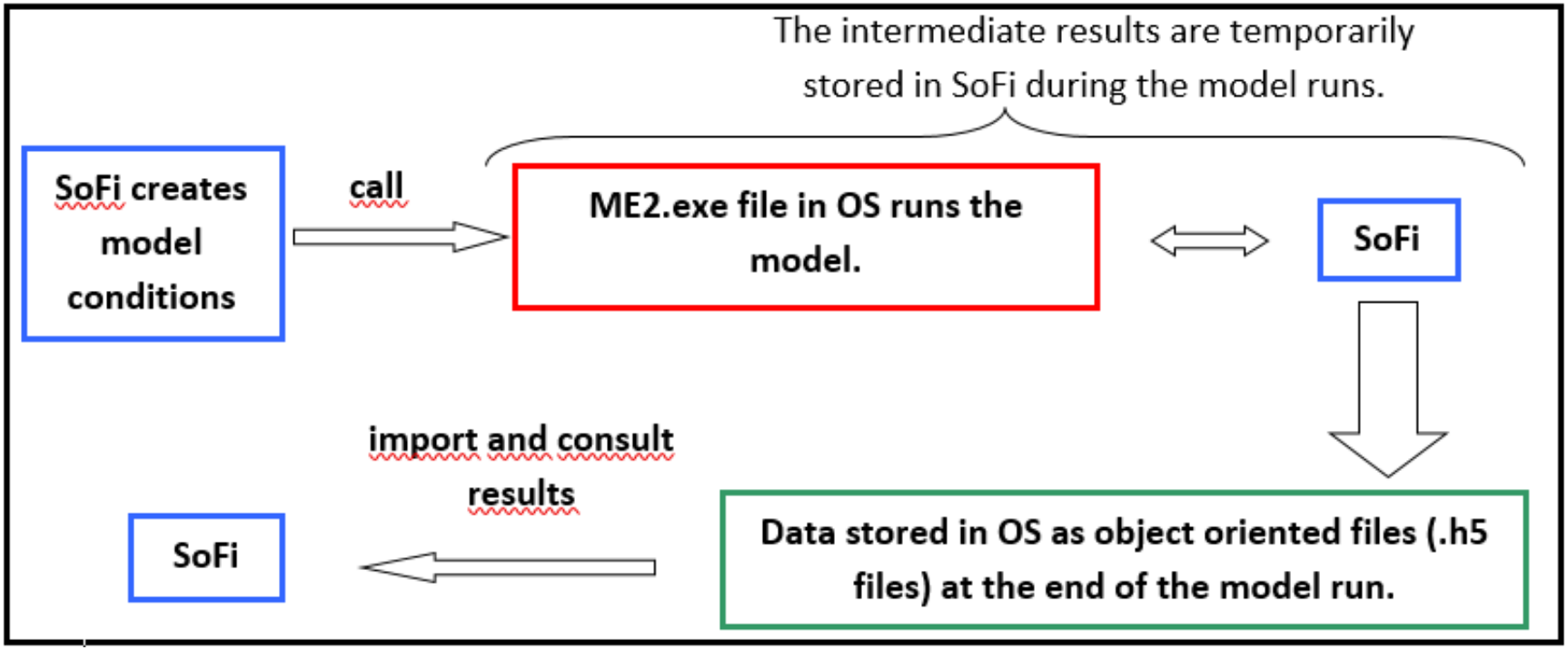
Pre-treatment of the input data

ME2.exe

Define setting for PMF run

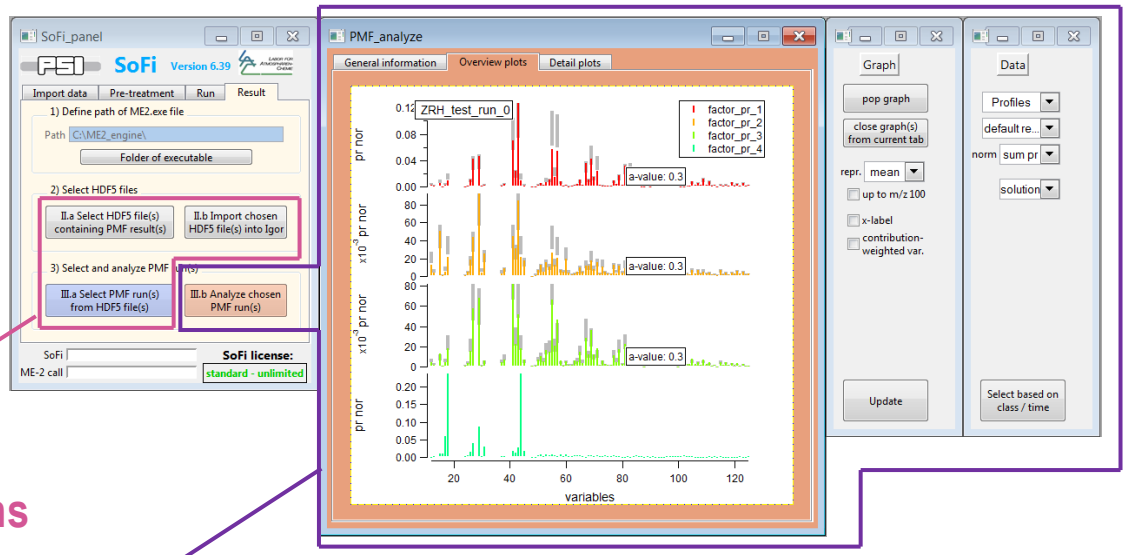
- General (#factors, missing data, etc.)
- Rotational ambiguity (seed, a-value, fpeak, pulling value)
- Statistical error propagation (add-on, released soon)
- Rolling mechanism (add-on, released soon)

SoFi/ME2 interaction



SoFi 6.39 – analyzing PMF results

ME2.exe



Select and import PMF runs for analysis

Tools for analyzing selected PMF runs

- Consult results quickly in preview window
  - General overview (time series, profiles)
  - Detail plots for time series and profiles
    - Fraction plots
    - Scatter plots
    - Correlation
    - (HR-family for HR-data)

PMF – some useful ressources

- Francesco Canonaco @ Paul Scherrer Institute : francesco.canonaco@psi.ch
- SoFi Google Group : sofi_me2@googlegroups.com
- Website : https://www.psi.ch/acsm-stations/me-2

- Source Apportionment Guide :
  European guide on air pollution source apportionment with receptor models (2013)

Useful Ressources

- Paatero's papers : (1994, 1999, 2009) on Positive Matrix Factorization and Multi-Linear Engine Model
- Canonaco's paper : (2013), on SoFi's toolkit http://www.atmos-meas-tech.net/6/3649/2013/
- Zhang's paper : (2011), Review on PMF https://link.springer.com/article/10.1007/s00216-011-5355-y
- Wiki page by Jimenez Group (for AMS data set) :
  - http://cires1.colorado.edu/jimenez-group/wiki/index.php/PMF-AMS_Analysis_Guide
  - AMS Spectral Database (HR) : http://cires1.colorado.edu/jimenez-group/HRAMSsd/
  - AMS Spectral Database (UMR) : http://cires1.colorado.edu/jimenez-group/AMSsd/

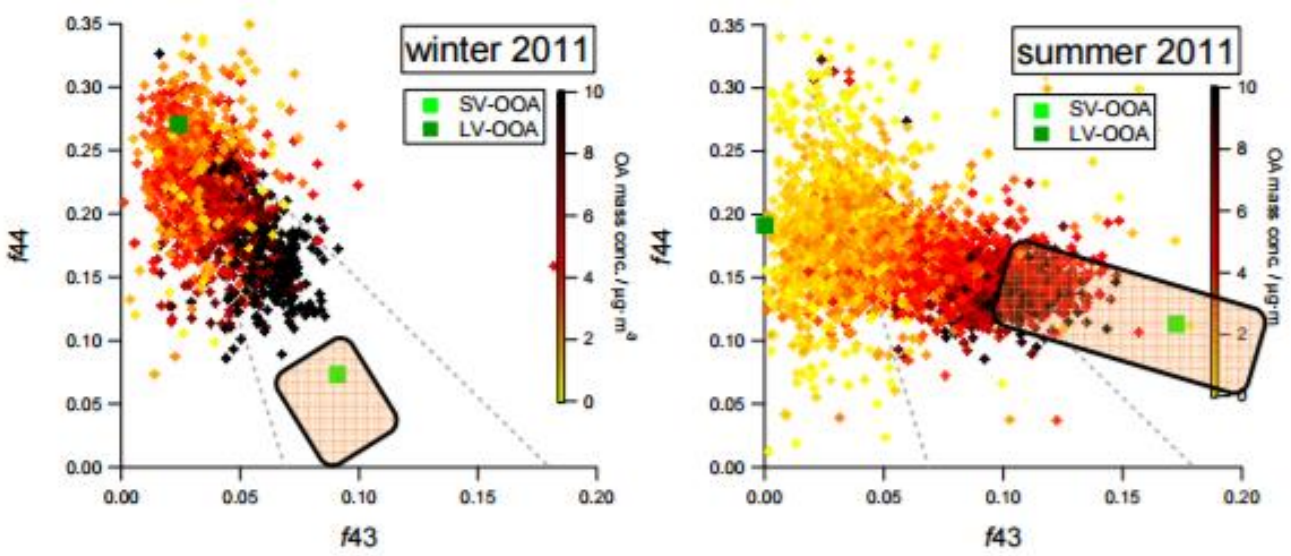
Prague, 15. - 17.2.2017

SoFi Pro

Key words:
Automated criteria-based selection, rotational & statistical uncertainty, C-value approach, additional variable / time separation

Variability of factor profiles

- limitation of PMF: factor profile is constant over the PMF run
- Example: Zurich ACSM data 2011/2012

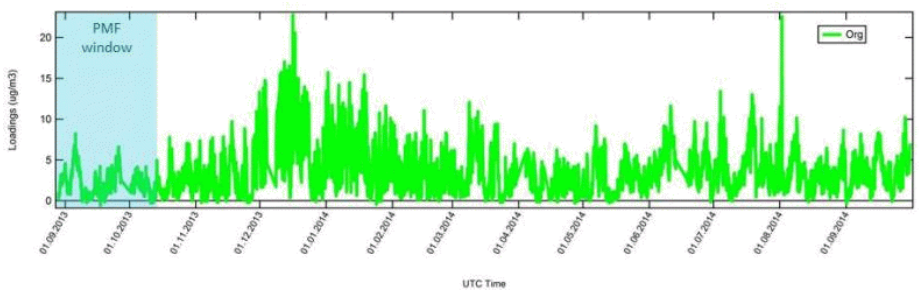


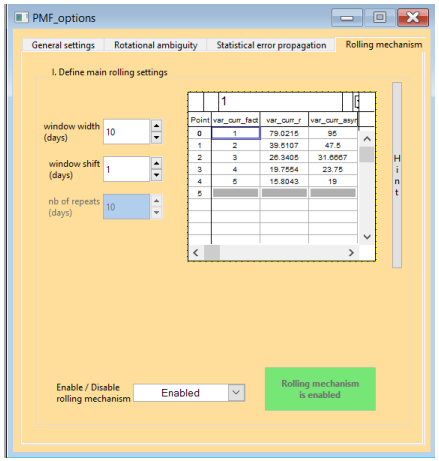
Canonaco et al. 2015

Automatic rolling PMF window

- Rolling PMF algorithm to account for seasonal and/or meteorological variations in OA sources

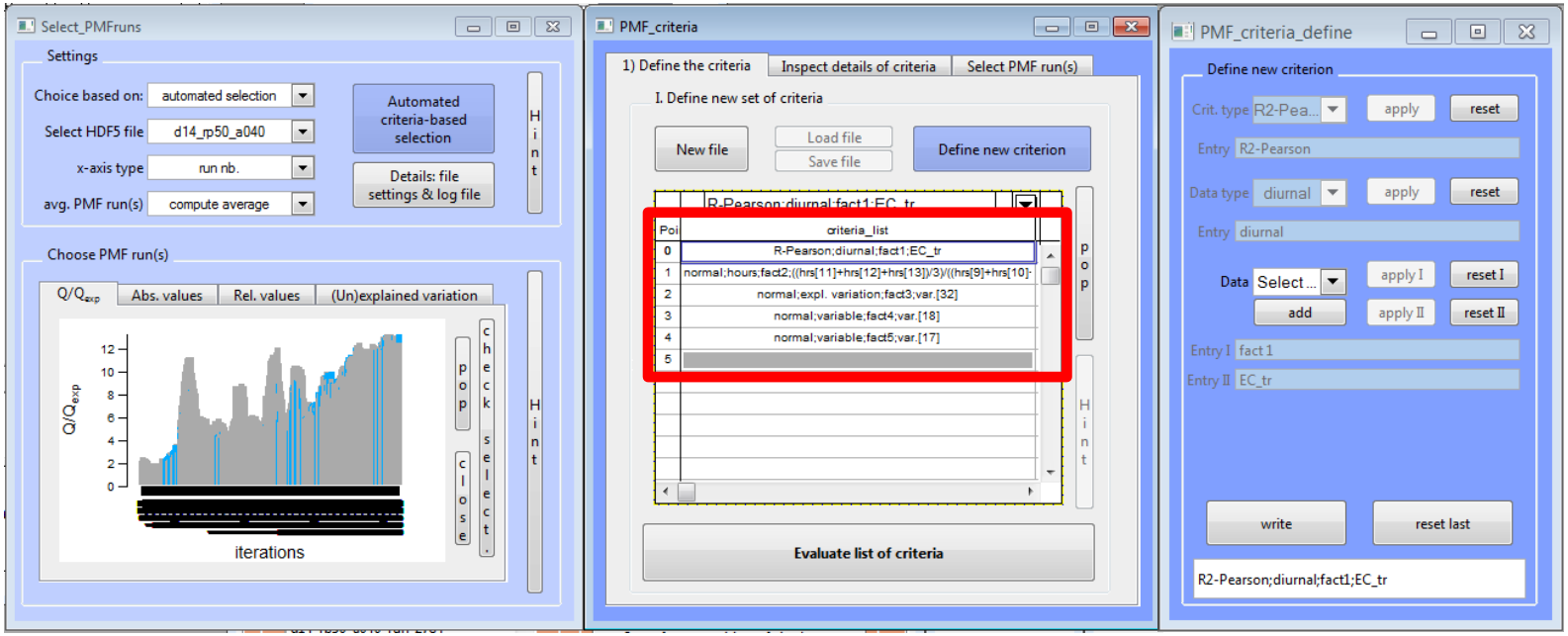
- PMF algorithm is run repeatedly on a short subset of data for a defined period
  - Assumption of constant aerosol sources during that time
  - after every shift the PMF runs are reinitialized (seed, *a*-value, fpeak, bootstrap, etc.)
- PMF window is subsequently shifted by defined period
- Thousands of runs that are sorted using the automated criteria-based selection
  - Goodness of PMF solution is estimated by using selection criteria such as correlation with external tracers (e.g. correlation of HOA with NO_x)





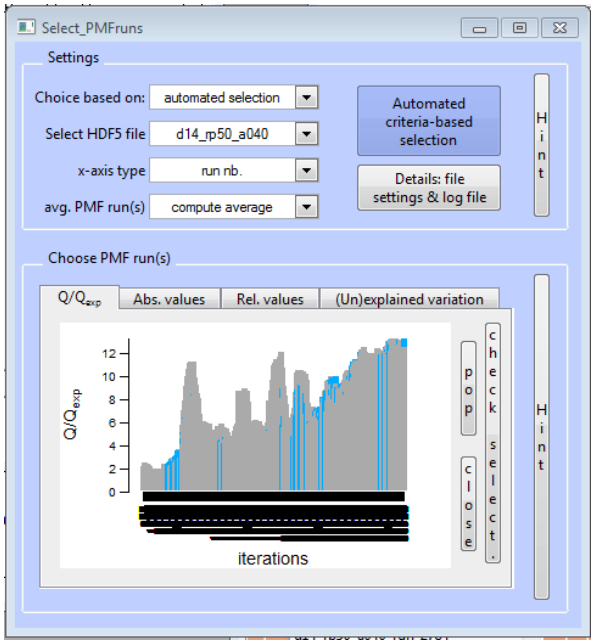
Automated criteria-based selection of PMF-runs

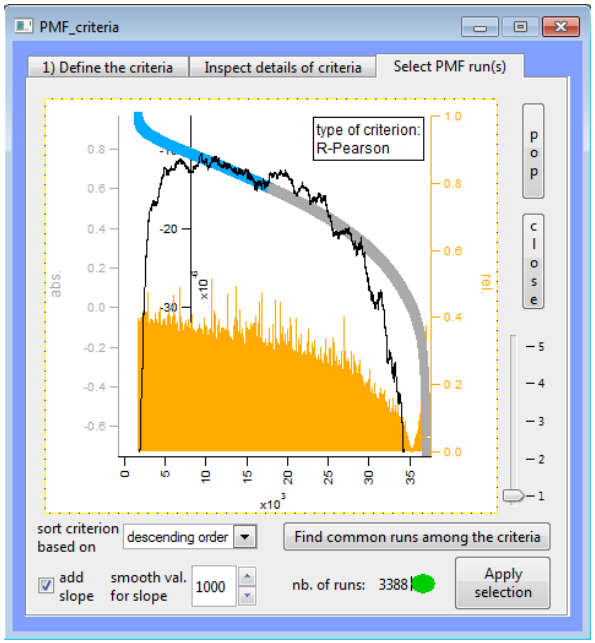
- User defines criteria for the factors, e.g. $R_{pearson}$ between contribution of factor 1 and NO_x (first entry)



Automated criteria-based selection of PMF-runs

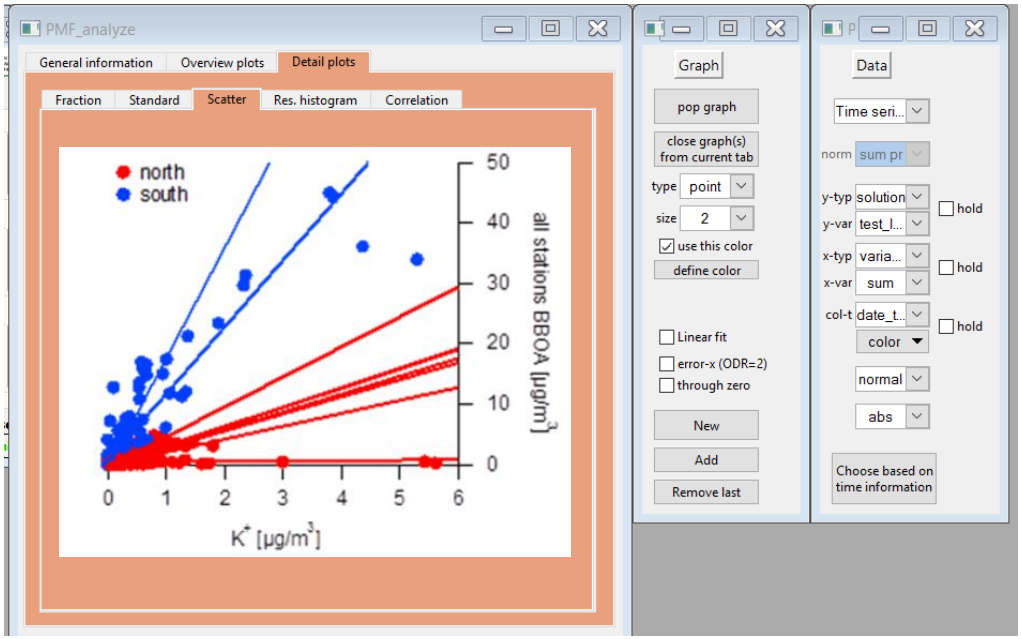
- all PMF runs are temporarily imported in IGOR and the scores of the criteria are evaluated for all runs. (example $R_{pearson}$, gray line)
- scores are sorted and highest ones selected for every criterion
- overlap is evaluated and PMF runs are selected for further analysis





Include additional information

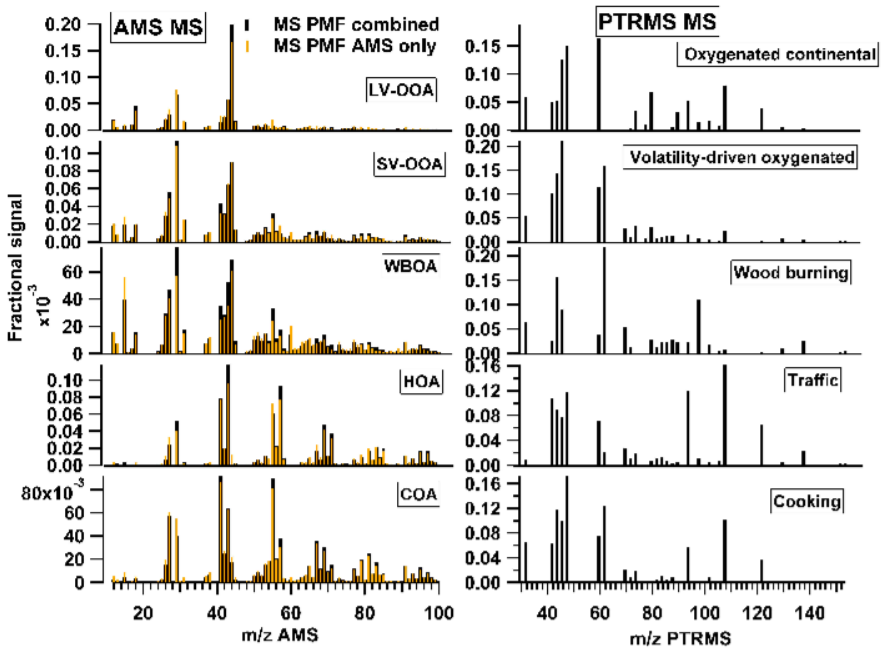
- Inspect PMF result based on additional information over variables / time
  - e.g. PMF run over data from two groups of data (north, south of the alps)

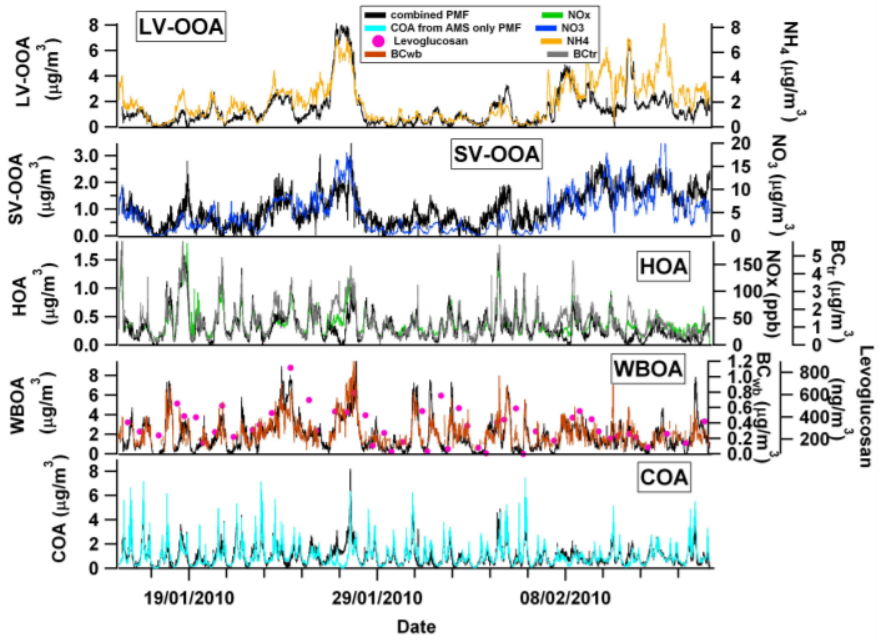


Daellenbach et al. in prep.

C-value approach

- automated weight of errors, e.g. when combining AMS with PTR-MS data





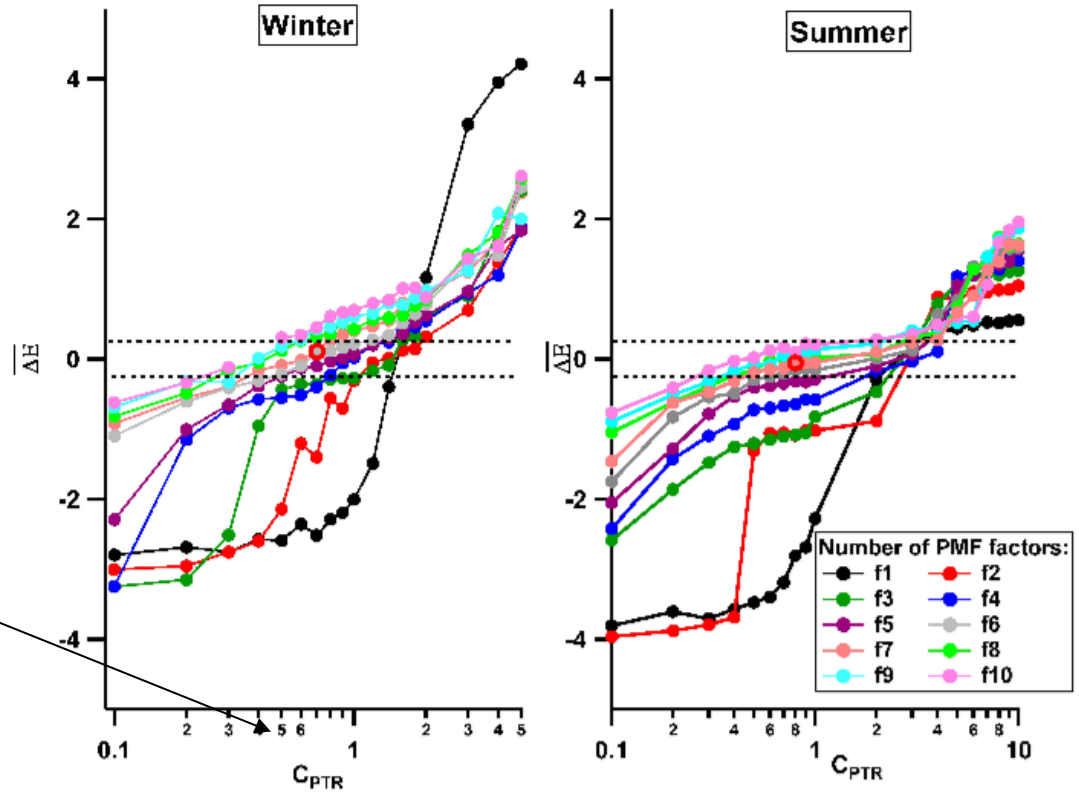
Crippa et al. 2013

C-value approach

- automated weight of errors, e.g. when combining AMS with PTR-MS data

$$\overline{\Delta E} = \overline{\left(\frac{|e_{ij}|}{s_{ij}}\right)}_{\text{AMS}} - \overline{\left(\frac{|e_{ij}|}{s_{ij}}\right)}_{\text{PTRMS}}$$

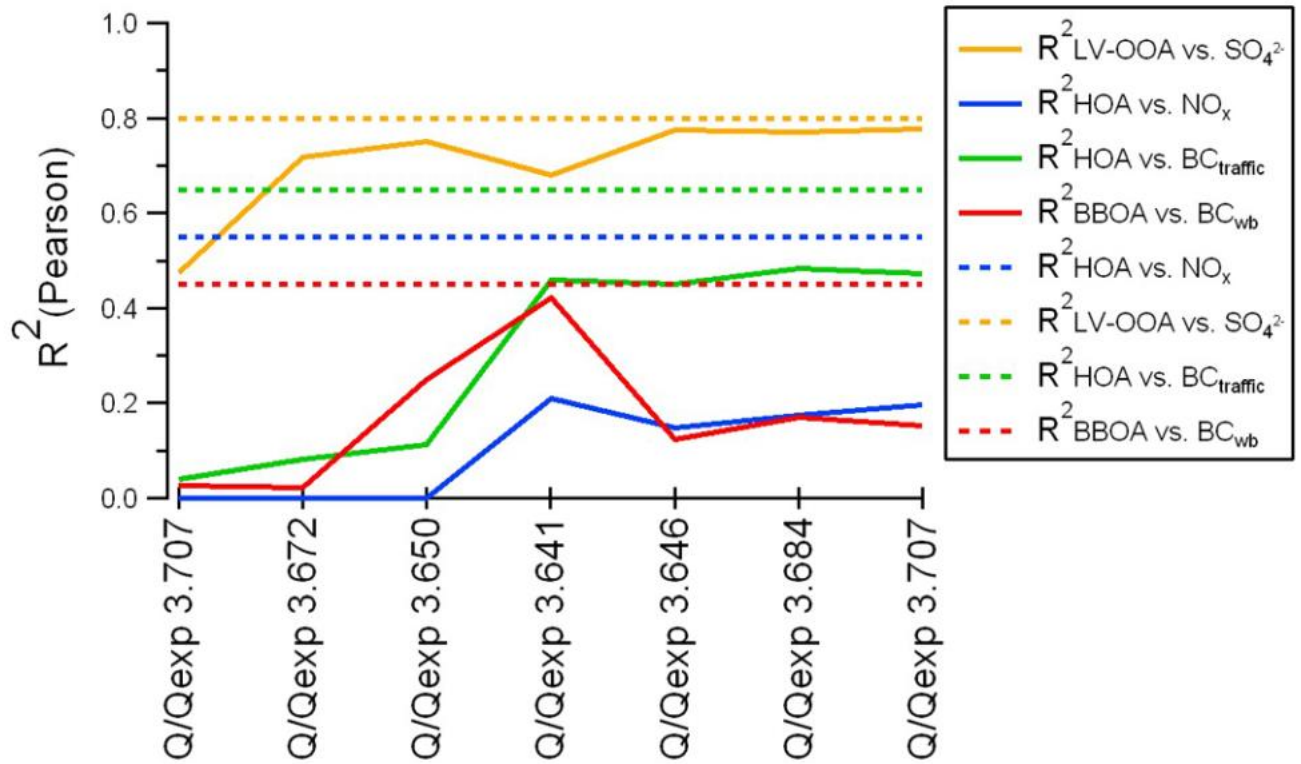
$$s_{ij,\text{new}} = \frac{s_{ij}}{C_{\text{PTR}}}.$$



Crippa et al. 2013

f-peak vs. a-value approach

- P. Paatero and PSI do **recommend to use the a-value approach**



PMF – 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

- example: three factors, rotation matrix **T** mixes all factor contributions and profiles together

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

Paatero 2008

PMF – 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: three factors, rotation matrix **T** mixes only factor 3 with factor 1 *(adding contribution of factor 3 to that of factor 1 and subtracting profile of factor 3 from 1)*

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

Paatero 2008