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.
Bilinear Model
Assumptions: Factors profiles (MS) are constant, Factor contributions (TS) vary
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)
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
* NOTE: All of the example diagnostics and graphs shown in this presentation are from PET
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 $\sum_i \text{Resid}_i^2$

Regular Least-Square minimizes $\text{Chi}^2 = \sum \text{Resid}_i^2$
PMF Approach: Least Squares Fit in multiple dimensions

\[ X_{ij} = \sum_p G_{ip} F_{pj} + \text{Resid}_{ij} \]

Data matrix

\[
\begin{array}{cccc}
X_{11} & X_{12} & X_{13} & \cdot & \cdot & \cdot \\
X_{21} & \cdot & \cdot & \cdot & \cdot & \cdot \\
X_{31} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]

Error matrix

\[
\begin{array}{cccc}
\sigma_{11} & \sigma_{12} & \sigma_{13} & \cdot & \cdot & \cdot \\
\sigma_{21} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\sigma_{31} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]

Q matrix

\[
\begin{array}{cccc}
Q_{11} & Q_{12} & Q_{13} & \cdot & \cdot & \cdot \\
Q_{21} & \cdot & \cdot & \cdot & \cdot & \cdot \\
Q_{31} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]

\[ Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \frac{\text{resid}_{ij}}{\sigma_{ij}} \right)^2 \]

estimated error of each point

Minimized Q during fit

Points with high Q drive the fit (i.e. high residuals relative to \( \sigma_{ij} \))

Data points weighted by \( \sigma_{ij} \)

Generated from AMS/ACSM Data matrices
For Good Fit, $Q_{\text{exp}} \approx m \times n$

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

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

For Good Fit, $Q/Q_{\text{exp}} \approx 1$
Noise Estimates for PMF

• Poisson Statistics (Counting Noise)

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

- \( \Delta I_d \) = 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

Not included

• m/z calibration shifts, peak width uncertainties

• peak fitting errors (Effect of neighbor intensity and spectral distance on fitting)
**Downweighting and Discarding**

\[ X_{ij} = \sum_p G_{ip} F_{pj} + \text{Resid}_{ij} \]

Q matrix

<table>
<thead>
<tr>
<th></th>
<th>Q_{11}</th>
<th>Q_{12}</th>
<th>Q_{13}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_{21}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q_{31}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \frac{\text{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 N₂ 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$_2$SO$_4$ (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 with 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

- Large $Q/Q_{\text{exp}}$ for some points (or mzs)
- $Q/Q_{\text{exp}}>1$ for most of dataset.
  - Underestimate Errors?
- Dataset doesn’t follow PMF model assumptions?

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

estimated error of each point

What features are driving the fit?
PMF Diagnostics

$X_{ij} = \sum_p G_{ip} F_{pj} + \text{Resid}_{ij}$

Scaled Residual $ij = \frac{\text{Resid}_{ij}}{\sigma_{ij}}$

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

Plateau Region in Q/Qexp

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

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

Any of these pairs could be used.

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

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


Introduction to Igor software programs

PET:

SOFI:
References

Other papers describing PMF of AMS data


PMF of Mixed AMS/PTRMS dataset

Questions?

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