Source Finder (SoFi)

Manual for the software package SoFi (Pro) in IGOR Wavemetrics Inc.

written by
Datalystica Ltd.
at Park InnovAARE
CH-5232 Villigen, Switzerland

last formatted: 12/14/2020
Manual for the Version of the software: SoFi 8.00 beta
Table of contents

1. General considerations ................................................................. 7
   1.1. What is SoFi ........................................................................ 7
   1.2. For whom is it meant? ......................................................... 7
   1.3. What is required? ............................................................... 7
   1.4. More details about the structure of SoFi? ......................... 8
   1.5. Differences between the SoFi versions? ......................... 8
2. PMF theory ................................................................................. 9
   2.1. The bilinear model – general considerations .................. 9
       2.1.1. General bilinear equation ........................................... 9
       2.1.2. Robust mode ......................................................... 10
   2.2. Normalizing Q by the expected value of Q (Q_{exp}) ........ 10
   2.3. Rotational ambiguity of the model solution ..................... 11
   2.4. Rotational techniques in ME-2 .......................................... 11
       2.4.1. Seed runs ............................................................ 11
       2.4.2. Global fpeak ...................................................... 11
       2.4.3. Individual fpeak .................................................. 12
       2.4.4. Partially constrained matrix F and/or G (α-value approach) .................................................. 12
       2.4.5. Partially constrained matrix F and/or G (pulling approach) .................................................. 13
3. Overview .................................................................................. 14
   3.1. Installation of the HDF5 interface ..................................... 14
   3.2. General structure .............................................................. 14
   3.3. First steps .......................................................................... 15
       3.3.1. Initialize SoFi ......................................................... 15
       3.3.2. Recreate panel, i.e.... .......................................... 15
       3.3.3. Recreate settings ................................................. 15
       3.3.4. Recreate SoFi ..................................................... 15
       3.3.5. Change main settings ............................................ 16
           3.3.5.1. Signal to noise (S/N) ...................................... 16
           3.3.5.2. Downweight values for results ...................... 17
           3.3.5.3. Robust values for result ................................ 17
           3.3.5.4. S/N threshold for the unexplained variation .................. 17
           3.3.5.5. Resolution in daily cycle ............................. 18
           3.3.5.6. Customize colors for the factor position .......... 18
       3.3.6. Manual ................................................................. 18
       3.3.7. User agreement ...................................................... 18
3.3.8. Credits............................................................................................................................ 18
3.4. SoFi Pro license activation step ..................................................................................... 19
3.5. Data hierarchy in IGOR .................................................................................................... 20
4. Import and define the input data .................................................................................... 21
4.1. Import data ...................................................................................................................... 21
  4.1.1. PMF Import .................................................................................................................. 22
  4.1.2. Time series import ...................................................................................................... 23
  4.1.3. Profile series import .................................................................................................. 24
4.2. Define PMF input ............................................................................................................ 25
  4.2.1. The data and error matrix ........................................................................................ 25
  4.2.2. The time wave ........................................................................................................... 25
  4.2.3. The variable vectors .................................................................................................. 25
  4.2.4. Data type ................................................................................................................... 25
    4.2.4.1. Minimum error for AMS data ............................................................................. 26
5. SoFi – pre-treatment .......................................................................................................... 27
5.1. Treat missing data in PMF input .................................................................................... 27
5.2. Treat external data ......................................................................................................... 28
  5.2.1. Profiles ....................................................................................................................... 28
  5.2.2. Time series ................................................................................................................ 29
    5.2.2.1. Wind data ............................................................................................................. 30
5.3. Explore PMF input ......................................................................................................... 30
  5.3.1. Options for the variable(s) ....................................................................................... 30
  5.3.2. Select based on class / time .................................................................................... 31
  5.3.3. Options for the graph representation ..................................................................... 32
5.4. Prepare PMF input for the run ..................................................................................... 34
  5.4.1. Select or modify effective PMF input ...................................................................... 34
    5.4.1.1. Class(es) .............................................................................................................. 34
    5.4.1.2. Input data ............................................................................................................. 37
6. SoFi – PMF run(s) ............................................................................................................. 40
6.1. Define path, nb. of .exe and name of HDF5 file .......................................................... 40
  6.1.1. Path .......................................................................................................................... 40
  6.1.1. nb. of executables ...................................................................................................... 40
  6.1.2. HDF5 file name ......................................................................................................... 41
6.2. Define general model options ...................................................................................... 41
  6.2.1. Define number of factors ........................................................................................ 41
  6.2.2. Convergence criteria ............................................................................................... 41
  6.2.3. Missing data .............................................................................................................. 42
  6.2.4. Error matrix (general) ............................................................................................ 43
6.2.4.1. Linear Error ................................................................. 43
6.2.4.2. Relative scaling ....................................................... 44
6.2.1. Error matrix (weight) .................................................. 46
6.2.1. Other various options ............................................... 48
6.2.1.1. Robust model threshold (A) ................................. 48
6.2.1.2. Model normalization (B) ......................................... 48
6.2.1.1. Mass closure (C) .................................................... 48
6.2.1.1. Multi-time (D) ....................................................... 48

6.3. Define options for the model matrices .................................. 51
6.3.1. Seed ........................................................................... 51
6.3.1.1. Main settings ....................................................... 51
6.3.1.2. Options for “from vector (fv)” ............................... 52
6.3.2. fpeak ........................................................................ 54
6.3.2.1. Global fpeak .......................................................... 54
6.3.2.2. Individual fpeak .................................................... 55
6.3.3. α-value ...................................................................... 56
6.3.3.1. Main settings ........................................................ 57
6.3.3.2. Exact α-value (exact) ........................................... 57
6.3.3.3. 1st dim. sensitivity (1dim. sens) ............................. 57
6.3.3.3. 2nd dim. sensitivity (2dim. sens) to 5th dim. sensitivity (5dim. sens) .................................................. 58
6.3.3.3. Random α-value (random) .................................... 58
6.3.3.4. Pass limits directly to ME-2 (limits) ......................... 58
6.3.3.5. Constraining variables ........................................ 58
6.3.3.6. All model variables ............................................. 58
6.3.3.7. All model variables of external vector .................... 59
6.3.4. Pulling ........................................................................ 60
6.3.4.1. Main settings ....................................................... 61
6.3.4.2. Exact pulling (exact) ............................................ 61
6.3.4.2. 1st dim. sensitivity (1D) ....................................... 61
6.3.4.4. 2nd dim. sensitivity (2D) to 5th dim. sensitivity (5D) ................................................................. 61
6.3.4.5. Random pulling values (RND) ............................... 62
6.3.1. Factor specifications .................................................... 65
6.3.2. Consideration on the priority of the rotational tools ....... 66

6.4. Estimation of the statistical uncertainty ................................ 67
6.4.1. Resampling strategy „bootstrap“ ................................ 67
6.5. Rolling mechanism .............................................................. 69
6.6. Call ME-2 and perform PMF ............................................. 70

7. SoFi – PMF result(s) ............................................................ 71
7.1. Define path of ME2.exe file ............................................ 71
7.2. HDF5 file(s) ................................................................. 71
7.2.1. Select HDF5 file(s) containing the PMF run(s) ............. 71
7.2.2. Import the chosen HDF5 file(s) into IGOR ................... 72
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>Select and analyze PMF run(s)</td>
<td>72</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Select PMF run(s) from HDF5 file(s)</td>
<td>72</td>
</tr>
<tr>
<td>7.3.1.1</td>
<td>Settings</td>
<td>72</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Manual selection of PMF run(s) form a HDF file</td>
<td>73</td>
</tr>
<tr>
<td>7.3.3</td>
<td>Criteria-based selection</td>
<td>74</td>
</tr>
<tr>
<td>7.3.3.1</td>
<td>Type of criterion</td>
<td>75</td>
</tr>
<tr>
<td>7.3.3.2</td>
<td>Options to set up a criterion</td>
<td>76</td>
</tr>
<tr>
<td>7.3.3.3</td>
<td>Examples on how to set up the list of criteria</td>
<td>81</td>
</tr>
<tr>
<td>7.3.3.4</td>
<td>Evaluate the criteria</td>
<td>88</td>
</tr>
<tr>
<td>7.3.3.5</td>
<td>Inspect criteria score</td>
<td>88</td>
</tr>
<tr>
<td>7.3.3.6</td>
<td>Inspect selected PMF runs</td>
<td>92</td>
</tr>
<tr>
<td>7.3.3.7</td>
<td>Inspect criteria stability</td>
<td>93</td>
</tr>
<tr>
<td>7.3.3.8</td>
<td>Different number of factors in one HDF file</td>
<td>94</td>
</tr>
<tr>
<td>7.3.3.9</td>
<td>Bootstrap runs on the base case</td>
<td>96</td>
</tr>
<tr>
<td>7.3.1</td>
<td>File settings and logfile</td>
<td>99</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Analyze chosen PMF run(s)</td>
<td>99</td>
</tr>
<tr>
<td>7.3.2.1</td>
<td>Averaging</td>
<td>100</td>
</tr>
<tr>
<td>7.3.2.2</td>
<td>Averaging and data options</td>
<td>100</td>
</tr>
<tr>
<td>7.3.2.3</td>
<td>Factor specification</td>
<td>101</td>
</tr>
<tr>
<td>7.4</td>
<td>Information on the result folder</td>
<td>102</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Structure of the result folders</td>
<td>102</td>
</tr>
<tr>
<td>7.4.2</td>
<td>Parameter files</td>
<td>103</td>
</tr>
<tr>
<td>7.4.2.1</td>
<td>Solution para</td>
<td>103</td>
</tr>
<tr>
<td>7.4.2.2</td>
<td>Batch file</td>
<td>104</td>
</tr>
<tr>
<td>7.4.3</td>
<td>Main information</td>
<td>104</td>
</tr>
<tr>
<td>8</td>
<td>SoFi – PMF run(s) analysis</td>
<td>106</td>
</tr>
<tr>
<td>8.1</td>
<td>Select a PMF run</td>
<td>106</td>
</tr>
<tr>
<td>8.2</td>
<td>General information</td>
<td>107</td>
</tr>
<tr>
<td>8.2.1</td>
<td>Select PMF run for exploration</td>
<td>107</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Info for all PMF run(s) in adjacent list</td>
<td>107</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Settings for chosen PMF run from adjacent list</td>
<td>107</td>
</tr>
<tr>
<td>8.2.4</td>
<td>Extract PMF solution</td>
<td>107</td>
</tr>
<tr>
<td>8.3</td>
<td>Subpanels graph and data</td>
<td>108</td>
</tr>
<tr>
<td>8.4</td>
<td>Overview plots</td>
<td>108</td>
</tr>
<tr>
<td>8.5</td>
<td>Detail plots</td>
<td>110</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Fraction</td>
<td>110</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Standard</td>
<td>111</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Scatter</td>
<td>113</td>
</tr>
<tr>
<td>8.5.3.1</td>
<td>Option model subtraction</td>
<td>114</td>
</tr>
<tr>
<td>8.5.4</td>
<td>Residuals</td>
<td>116</td>
</tr>
<tr>
<td>8.5.4.1</td>
<td>Entire residual matrix</td>
<td>116</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>8.5.4.2</td>
<td>Residual histogram</td>
<td>116</td>
</tr>
<tr>
<td>8.5.5</td>
<td>Correlations</td>
<td>119</td>
</tr>
<tr>
<td>8.5.6</td>
<td>HR plot</td>
<td>120</td>
</tr>
<tr>
<td>8.6</td>
<td>Averaged solution</td>
<td>121</td>
</tr>
<tr>
<td>8.6.1</td>
<td>Repeats</td>
<td>121</td>
</tr>
<tr>
<td>8.6.1</td>
<td>Mass error estimation</td>
<td>122</td>
</tr>
<tr>
<td>8.6.2</td>
<td>Employed $\sigma$-values</td>
<td>123</td>
</tr>
<tr>
<td>8.6.1</td>
<td>t-profiles</td>
<td>124</td>
</tr>
<tr>
<td>9</td>
<td>List of definitions and abbreviations</td>
<td>126</td>
</tr>
<tr>
<td>10</td>
<td>Acknowledgment</td>
<td>127</td>
</tr>
<tr>
<td>11</td>
<td>Literature</td>
<td>128</td>
</tr>
</tbody>
</table>
1. General considerations

1.1. What is SoFi

The software Source Finder (SoFi) for initiating and controlling the multilinear engine algorithm (ME-2, P. Paatero, U. of Helsinki, Finland) that solves the positive matrix factorization algorithm (PMF, P. Paatero, U. of Helsinki, Finland), as well as analyzing the ME-2 results was written at the Paul Scherrer Institute (PSI) and Datalystica Ltd. in Villigen, Switzerland. SoFi is a finite software package that runs within the software environment IGOR Pro (Wavemetrics Inc., Lake Oswego, OR, USA).

1.2. For whom is it meant?

SoFi is intended to facilitate the source apportionment of measured ambient data. The software was originally developed for analysis of aerosol chemical speciation monitor (ACSM) data (Ng et al., 2011). However, the new software with its highly panel-oriented structure guarantees a seamless source apportionment analysis from the data input to the data output with basically no interaction at all with the software environment IGOR making SoFi fully suitable even for data analysts with no programming or knowledge skills of IGOR. The algorithm can be applied to any kind of data, for which dimensionality reduction analysis based on the co-variance of the variables of interest, is aimed and will be distributed to any interested data analysts under the terms and conditions of Datalystica, the official distributor of SoFi (consult the homepage of Datalystica (https://datalystica.com) for the terms and conditions). The same terms and conditions are visualized when SoFi is launched and the user has to read and agree to them, before being able to use the SoFi software.

1.3. What is required?

Make sure you already possess an official key for ME-2 (must be purchased directly from Dr. P. Paatero, email to pentti.paatero86@gmail.com or from the homepage of Datalystica, https://datalystica.com), full administrator rights of your PC, as SoFi will create, modify and execute script files on the operating system and an official license for the software IGOR 6 or 8 (must be purchased directly from Wavemetrics, https://www.wavemetrics.com or very soon also directly from the homepage of Datalystica, https://datalystica.com).
1.4. More details about the structure of SoFi?

SoFi consists of a standard version (freeware and unlimited license) and SoFi Pro (commercial key that enables all advanced functionalities of SoFi). SoFi Pro key can be purchased from Datalystica, [https://datalystica.com](https://datalystica.com). Consult the homepage of Datalystica for more details. This manuscript covers both, all functionalities of the SoFi software for the standard but also for the Pro version. Utilities that are only available when enabled by the SoFi Pro key are written in green to better separate them from the standard utilities of SoFi.

This manuscript does not provide a detailed description of the theory behind the PMF algorithm nor on the ME-2 solver and its script code but rather the operation of the abovementioned software. Nonetheless, a small general theoretical introduction on PMF is presented in the next chapter. Interested readers are referred to the original paper of the multilinear engine (Paatero, 1999), the papers introducing the positive matrix factorization algorithm (Paatero and Tapper, 1994; Paatero, 1997; Paatero, 2004), the paper that provides an overview over the different rotational tools of the ME-2 algorithm (Paatero and Hopke, 2009) and the recent methodological papers about ME-2 (Brown et al., 2015; Paatero et al., 2014). In addition, the reader is also referred to Paatero (Paatero, 1997) for a discussion on the robust mode and on (Paatero and Hopke, 2003) for a detailed discussion on the reweighting strategy.

The reader is also referred to Canonaco et al., (2013) and (2015) for the introduction paper of SoFi and its application on seasonal ACSM data, respectively as well as to Crippa et al., (2014), where the $\alpha$-value approach was thoroughly tested for many European AMS case studies, averaged profiles for the primary organic aerosol (POA) were retrieved and a general strategy using the $\alpha$-value approach for AMS data is presented and discussed.

1.5. Differences between the SoFi versions?

There are currently two main versions of SoFi that are operational in IGOR, i.e., SoFi 6 for IGOR 6 and SoFi 8 for IGOR 8. The description of the features of SoFi contained in this manual refers to the latest version, i.e., SoFi 8. Note that many of the features presented here are also available for SoFi 6. However, SoFi 6 will no longer be updated, only major bugs will be resolved therein. Moreover, it is not guaranteed that SoFi 6 for IGOR 6 and SoFi 8 for IGOR 8 will run completely error-free all the time. There might be cases and conditions that have not been tested or encountered during our test-phases, and that might cause some unexpected problems on SoFi or IGOR. We apologize for this eventuality and strongly rely on the bug report of all SoFi users to minimize such eventualities but also for improving SoFi for your needs! Thanks already in advance for your feedback! Any feedback, no matter what, is always highly appreciated!

Your Datalystica Team!
2. PMF theory

2.1. The bilinear model – general considerations

2.1.1. General bilinear equation

The measurement of a set of variables (e.g. UMR m/z intensities for the ACSM) as a function of time can be represented as a data matrix $X$ where the columns $j$ are the variables and each row $i$ represents one single scan.

In so-called factor analytical models, where a dimension reduction occurs, the variables are grouped in a few factors that can be further investigated rather than dealing with all measured variables. A commonly used approach is to group variables into two matrices using the so-called bilinear model. Few examples of bilinear models are the principal component analysis PCA (Wold et al., 1987), chemical mass balance approach (Miller et al., 1972) or the positive matrix factorization PMF (Paatero and Tapper, 1994). A bilinear factor model in matrix notation is defined as:

$$ X = GF + E $$ \hspace{1cm} (1)

where the measured matrix $X$ is approximated by the product of matrices $G$ and $F$. $p$, i.e. the number of columns of the modelled matrix $G$ and the number of rows of the modelled matrix $F$ is called the number of factors. Each column $j$ of the matrix $G$ represents the time series of each of the $p$ factors, whereas each row $i$ of $F$ represents the corresponding factor profile or fingerprint (e.g. mass spectrum). Note that by definition of (1) factor profiles are static and are not allowed to vary over time. The matrix $E$ represents the model residual. The differences between PCA and PMF are mainly due to the restrictions of these models. PCA imposes orthogonality on the factor profiles, i.e., the scalar of two different rows of $F$ is zero and does not require non-negative entries. In contrast, PMF require non-negative entries throughout $G$ and $F$. PMF2 and ME-2 are solvers capable of solving the PMF algorithm. Both solvers fit the entries in $G$ and $F$ using a least squares algorithm that minimizes iteratively the quantity $Q_m$, i.e., the total sum of the squared model residuals $e_{ij}$ weighted by the uncertainty $\sigma_{ij}$ for all input points $ij$ (2):

$$ Q_m = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{e_{ij}^2}{\sigma_{ij}^2} \hspace{1cm} (2) $$

Here, the uncertainties $\sigma_{ij}$ typically contain the measurement uncertainty for the input point $ij$. Data points for which the ratio $e_{ij} / \sigma_{ij} \gg 1$ constitute a large fraction of $Q_m$ and consequently these points will have a high impact during the model iteration. Normally this condition ensures that data with high signal-to-noise ratio is considered more than data from measurements close
to the detection limit. However, for some data points a case when the ratio \( e_{ij} / \sigma_{ij} \gg 1 \) may also occur due to dominant local events or electronic noise within the measurement equipment, neither of which should be considered by the model.

### 2.1.2. Robust mode

To prevent a solution to be driven by these events, the model is generally run in a “robust” mode, in which pulling of the solution by outliers is reduced. At each step of the iteration process, outliers are defined based on the ratio of residuals to uncertainties:

\[
\text{outlier} := \frac{e_{ij}}{\sigma_{ij}} > \alpha \tag{3}
\]

\( \alpha \) is the user-defined threshold value. A value of four is recommended as a defining criterion for outliers within the robust mode (Paatero, 1997). The residuals are reweighted dynamically to remove the dependence of the change of \( Q^m \) with respect to the change of the outlier residuals.

\[
\frac{dQ^m}{dE_{\text{outliers}}} = \text{const.} \tag{4}
\]

This prevents the model solution to be driven by few dominant points in the datasets. It is recommended to always run the model in the “robust” mode, when treating data prone to e.g. rare transient sources, spikes, electronic noise or low signal to noise ratio, typically the case for ambient data.

### 2.2. Normalizing Q by the expected value of Q (\( Q_{\text{exp}} \))

Normally, monitoring and comparing total \( Q \) among various PMF runs is not meaningful because the expected value depends on the size of the data matrix and on the number of chosen factors. One therefore normalizes \( Q^m \) by the remaining degrees of freedom that is both a function of the size of the data matrix and of the number of factors and is called \( Q_{\text{exp}} \).

\[
Q_{\text{exp}} = n \cdot m - p \cdot (n + m) \tag{5}
\]

In past, AMS studies reported typical values for the ratio of \( Q/Q_{\text{exp}} \) between 1 and 5. Sometimes values below 1 were reported. Note that when running PMF in the robust mode with a robust mode value of 4 to highest value of \( Q/Q_{\text{exp}} \) is ~16. For an ideal PMF run the \( Q/Q_{\text{exp}} \) would decrease down to ~1 only if the employed uncertainty considers both the measured as well as the modeled uncertainty. The latter is very difficult to assess and it remains often unknown, especially for ambient data. In addition, downweighting procedures (see 6.2.1) may have an impact on the value of \( Q \), depending on the amount and magnitude of weak / bad variables. Therefore, the absolute value cannot be strictly used as a metric for judging model results. Instead, one should
investigate the relative change of \( Q/Q_{\text{exp}} \) across different model runs as guidance in choosing reasonable model solutions.

2.3. Rotational ambiguity of the model solution

General bilinear models possess a huge number of solutions, as \( \mathbf{G} \) and \( \mathbf{F} \) can be transformed with \( \mathbf{T} \) and \( \mathbf{T}^{-1} \), respectively (see equation (6)). Bilinear models with restrictions as orthogonality or non-negativity decrease this number of rotations (also called degree of rotational ambiguity), but it remains high (Paatero et al., 2002). There are two different kinds of rotations that are allowed occur, the pure and the approximate rotations. For pure rotations, the object function \( Q^m \) stays invariant with respect to rotations in space:

\[
\mathbf{\tilde{G}} = \mathbf{GT} \quad \text{and} \quad \mathbf{\tilde{F}} = \mathbf{T}^{-1}\mathbf{F}
\]

(6)

where \( \mathbf{T} \) is a nonsingular matrix of dimension \( p \times p \), \( \mathbf{T}^{-1} \) is its inverse and \( \mathbf{\tilde{G}} \) and \( \mathbf{\tilde{F}} \) are the rotated matrices. The matrix multiplication of \( \mathbf{\tilde{G}} \) and \( \mathbf{\tilde{F}} \) leads to the same product as for multiplication of \( \mathbf{G} \) and \( \mathbf{F} \), and therefore \( Q^m \) remains unchanged. If the transformation matrix \( \mathbf{T} \) does not fulfill equation (6), the rotation is called an approximate rotation, in case of which \( Q^m \) changes.

2.4. Rotational techniques in ME-2

2.4.1. Seed runs

Initializing the PMF model matrices with random or pseudo random data is called a seed run, as the model matrices are sowed / populated with data. This is typically performed when looking for the global minimum in the Q-space. However, as seed runs have various starting points, they will also fall at different locations at the global minimum, given that the global minimum is very likely to occur and the data has a high degree of rotational ambiguity. Therefore, seed runs also allow for a partial exploration of the rotational ambiguity. However, this technique does not allow for a systematic exploration and it is recommended for this purpose to use one of the techniques described below.

2.4.2. Global fpeak

For the PMF2 and the ME-2 engine, there is a user-specific parameter called fpeak, denoted by \( \phi \) for the global control of such rotations. For positive \( \phi \), elementary rotations or a series of elementary rotations are performed that increase columns of the matrix \( \mathbf{G} \) and decrease rows of the matrix \( \mathbf{F} \) while conserving mass. The opposite occurs for negative \( \phi \). However, the fpeak tool explores only rotations in one dimension of the multidimensional space and if the entries of \( \mathbf{G} \) and \( \mathbf{F} \) are positive and more than one factor is chosen then the rotational space is
multidimensional and the corresponding ambiguity can be very large (e.g. for three factors, the rotational space is nine-dimensional).

**For PMF2 users:** Note that the fpeak value in ME-2 is proportional to the fpeak of PMF2. They are not necessarily equal (Paatero and Hopke, 2009).

### 2.4.3. Individual fpeak

An advantage of the ME-2 solver compared to the PMF solver is improved rotational control, e.g. selected factors can be summed / subtracted together rather than transforming the entire matrix. Thus, the rotations can be studied in a more controlled environment. The analyst has to explore the solution space to provide the range of environmentally reasonable PMF runs upon the use of rotational tools. Rotational ambiguity can be very high, in particular for large datasets with many variables, long measurement times, high number of factors, low signal to noise ratios, etc. and hence, it is advantageous to introduce *a priori* information into the PMF model to reduce the amount of rotational ambiguity. This can happen in form of known rows of F (factor profiles) and/or of known columns of G (factor time series) (Paatero and Hopke, 2009). This *a priori* information prevents the model matrices to rotate, resulting in ~unique solutions. In ME-2 two different main approaches are implemented, the α-value and the pulling techniques which will be briefly discussed in the following subchapters.

### 2.4.4. Partially constrained matrix F and/or G (α-value approach)

In α-value approach the elements of the F matrix (factor profiles) or of the G matrix (factor time series) can be constrained by the user. The user inputs one or more factor profiles (rows of F) / factor time series (columns of G) and a constraint defined by the scalar α that can be applied to the entire profile / time series or to elements of the profile / time series only. The α-value determines the extent to which the output $F_{s\text{olution}} / G_{s\text{olution}}$ is allowed to vary from the input $F_{\text{initial}} / G_{\text{initial}}$, according to:

$$f_j, \text{solution} = f_j \pm \alpha \cdot f_j$$
$$g_i, \text{solution} = g_i \pm \alpha \cdot g_i$$

(7) (8)

where $f_j$ and $g_i$ represent the row and the column of the matrices F and G, respectively. The index $j$ varies between 0 and the number of variables and $i$ varies between 0 and the number of measured points.

The scalar α ranges between 0 and 1. Setting the α-value to zero for all the factor profiles, one can fully constrain the factor profile matrix F, similar to the method of a chemical mass balance (CMB). On the contrary, setting the α-value to one for all entries of a profile or time series allows it to similar as in an unconstrained situation.
2.4.5. Partially constrained matrix \( F \) and/or \( G \) (pulling approach)

The user has the possibility to introduce pulling equations into the model that pull profile factor elements towards target values, so-called anchors (here shown for the row \( j \) of the matrix \( F \) only):

\[
\overline{a}_j = f_j + r_j
\]  

(9)

In equation (9), \( \overline{a}_j \) represents the anchor to which the model has to pull the iterative value \( \overline{f}_{ij} \) and \( r_{ij} \) represents the residual. The anchor is some known value introduced as a priori information by the user. The pulling equations create an additional auxiliary term \( Q^{aux} \) that is added apart from \( Q^m \). Thus, if pulling equations are introduced, the model will minimize the argument of \( Q \) with respect to all entries in the matrices \( G \) and \( F \):

\[
\arg \min_{G,F} Q = \arg \min_{G,F} (Q^m + Q^{aux})
\]  

(10)

The term of \( Q^{aux} \) has a conceptually similar aspect to \( Q^m \):

\[
Q^{aux} = \sum_{k_i = 1}^{K} \left( \frac{r_k}{s_k} \right)^2
\]  

(11)

The index \( j \) from equation (9) has been replaced by \( k \), since \( k \) denotes the index of the pulling equations added to the model (over many factor profiles / time series). The pulling parameter \( s_k \) specifies the “softness” of the pull. The smaller \( s_k \) becomes, the higher the impact of \( Q^{aux} \) of the \( k \)-th pull during the iterative process. The pulling approach is a sensitive technique in that if the pulling equation is not compatible with the specific data matrix, i.e. a decrease of \( Q^{aux} \) obtained as \( \overline{f}_{ij} \) reaches its anchor value \( \overline{a}_j \) (9) will be negligible compared to a larger increase of \( Q^m \), then the pull is neglected. Generally, the user provides the acceptable limits of \( Q^m \) denoted as \( dQ \). Changing \( dQ \) and the pulling parameter \( s_k \) allows monitoring a change in \( Q^m \) and judging about its acceptability.

The anchor can be any entry of a factor profile / time series, a mathematical combination of one or more entries of a factor profile / time series and also involving external data.
3. Overview

3.1. Installation of the HDF5 interface

The storage of the model results is based on the object oriented HDF5 file system. Since the IGOR software does not load the procedures for the HDF5 file system by default, the user needs to link the required procedures manually in the operating system (OS).

Create a shortcut of the following files
I. C:\Program Files\WaveMetrics\Igor Pro Folder\More Extensions\File Loaders\HDF5.xop
II. C:\Program Files\WaveMetrics\Igor Pro Folder\More Extensions\File Loaders\HDF5 Help.ihf
III. C:\Program Files\WaveMetrics\Igor Pro Folder\WaveMetrics Procedures\File Input Output\HDF5 Browser.ipf

Place the shortcuts in
I. C:\Program Files\WaveMetrics\Igor Pro Folder\Igor Extensions\HDF5.xop
II. C:\Program Files\WaveMetrics\Igor Pro Folder\Igor Extensions\HDF5 Help.ihf
III. C:\Program Files\WaveMetrics\Igor Pro Folder\Igor Procedures\HDF5 Browser.ipf

3.2. General structure

The IGOR-based software handles all ME-2 implementations discussed in the previous chapter. The operational framework is presented in Figure 1.

![Diagram](image)

Figure 1 The general process from the model call to the storage of the model output. Blue rectangles describe SoFi tasks, red rectangle represents the ME-2 execution. After all model runs finished, the results are stored in an HDF5 file (green color).
3.3. First steps

The first step when initializing the SoFi software is to open a new IGOR experiment. Thereafter, the user needs to import the SoFi procedure file (double-click on the SoFi procedure file). The actual version has the name:

SoFi_standard_6.#.ipf (see first page for the current version)

Once the SoFi procedure is compiled, a new tab appears in the main menu in IGOR.

![Image of IGOR interface with new tab labeled SoFi]

Figure 2 The tab “SoFi” appears once the procedure is compiled.

The tab “SoFi” contains several main options:

3.3.1. Initialize SoFi

“Initialize SoFi” launches the SoFi code and sets all its variables and strings

Before launching SoFi the user must read and therefore scroll (mouse wheel, cursor or page down button) down the user’s agreement. Only after the cursor has reached the last line of the text the “agree and confirm..” button unfreezes and allows the user to continue.

3.3.2. Recreate panel, i.e....

“recreate panel, i.e....” closes all graphs and restarts the main panel. This button leaves the imported PMF results untouched.

3.3.3. Recreate settings...

“recreate settings...” additionally, resets all variables and strings related to SoFi. This button leaves the imported PMF results untouched.

3.3.4. Recreate SoFi......

This option is similar to “Initialize SoFi”. Everything will be recreated.
3.3.5. **Change main settings**

Allows the user to modify main settings. Currently, the following settings are available and modifiable from here: signal to noise used in SoFi (S/N), S/N threshold for the (un)explained variation plot, enable/disable the progress bar and colors of the PMF factor positions.

![Main settings panel for SoFi](image)

**Figure 3** Main settings panel for SoFi

### 3.3.5.1. **Signal to noise (S/N)**

The user defines which method of S/N calculation should be employed in SoFi for the cell-wise or averaged approach. This affects the variable S/N plot (see 5.3) or when estimating the (un)explained variation in the result plots (see 8.5). In case of activation of the cell-wise calculation, S/N ratios are estimated for each cell/entry of the PMF input separately, whereas the averaged treatment provides an averaged (over time) S/N ratio per variable.

Following options are available in SoFi:

**Cell-wise**

- \( \frac{\text{abs}(\text{data}_{ij})}{\text{error}_{ij}} \)
- \( \frac{\text{data}_{ij}}{\text{error}_{ij}} \)  

(Brown et al.,(2015)), see supplement in Brown et al.,(2015)

**Averaged**

- \( \frac{\text{sum}((\text{data}_i)^2)}{\text{sum}(\text{error}_i)^2}} \)
- \( \frac{\text{sum}(\text{data}_i)}{\text{sum}(\text{error}_i)} \)
- \( \text{sqrt}(\text{sum}((\text{data}_i)^2))/\text{sqrt}(\text{sum}(\text{error}_i)^2}) \)

(Brown et al.,(2015)), see supplement in Brown et al.,(2015)

“\( i \)” being the index running over time and “\( j \)” being the index for the variables.
3.3.5.2. **Downweight values for results**
The error matrix for a PMF result can either contain the downweighted values (default case) or be shown without the downweighted values (see 6.2.1). The former has the advantage that the user sees what the ME-2 solver tried to minimize during the iteration. The latter shows the residuals, without the downweighted values that is relevant (and gives also more realistic values for Q), in case many variables were classified as weak or bad and downweighted.

3.3.5.3. **Robust values for result**
The robust mode truncates the dependency of the iterative process from the scaled residuals (see 2.1.2), preventing e.g. outliers to drive a PMF solution, by rescaling and hence by modifying the corresponding errors. The user has the possibility to see the result including this modification (default case) or to remove this effect from the error matrix. The advantage of applying the robust mode is that the minimization process as solved by ME-2 can be inspected. Removing this option reveals which variables or points in time are better or worse explained among all points that hit and remained until convergence at the robust limit.

The last two points (3.3.5.2 and 3.3.5.3) can be performed while inspecting the PMF results. This allows the user to compare the result with and without the application of these threshold limits.

3.3.5.4. **S/N threshold for the unexplained variation**
Paatero (2010) introduced a metric based on the estimation of the measurement variation explained by the factors. The explained variation (EV) is a dimensionless quantity that indicates how much variation in time or variation in each variable is explained by each factor. As an example, the equation for the explained variation EV for the $i^{th}$ point in time for the $k^{th}$ factor is given by:

$$ EV_{ik} = \frac{\sum_{j=1}^{p} [g_{ik} \cdot f_{kj}] / \sigma_{ij}}{\sum_{j=1}^{p} (\sum_{h=1}^{m} [g_{ih} \cdot f_{kj}] + \epsilon_{ij}) / \sigma_{ij}} $$ (12)

Similar equations can be formulated for the unexplained variation (UEV) by replacing the product $g_{ik} \cdot f_{kj}$ in the numerator with $e_{ij}$. Expressing the explained and the unexplained variation for a variable $j$ as EV_{jk} is done by simply replacing in the expression the sum over $j$ with the sum over $i$. If all variation is explained by the model, then EV = 1. According to Paatero (2010) a variable should be regarded as explained only if the UEV for that variable is less than 25%.

The UEV is further separated into the real UEV for data possessing a high S/N value and UEV for noisy data. This separation is performed using the S/N threshold defined in this section (default value is 2).
3.3.5.5. Resolution in daily cycle
The resolution for the daily cycle plots supported in SoFi can be changed here. The default value is one that leads to the classical daily cycle resolution between 0 and 24. This option is interesting and necessary, in case the PMF input data or external data possess a lower resolution than one hour.

3.3.5.6. Customize colors for the factor position
The user has the possibility to define colors for the factor positions up to 100 positions. Pressing the “SHIFT” key selects adjacent factor positions. Pressing the “CTRL” key allows the user to add non-adjacent factor positions. Once the factors are chosen, pressing on “Start defining” pops the color palette, in which the colors for the chosen factor positions are defined. This process is additive, i.e. the user can define a second or third step other colors or existing factors or new colors for other factors.

![Figure 4](image)

**Figure 4** Options present for “change main settings” panel on the left. The right panel is used to select user-specific colors for the various factor positions.

3.3.6. Manual
“Manual” starts the SoFi manual, given that it is placed in the same folder as the ME2.exe file and the path to this file is selected in SoFi.

3.3.7. User agreement
“User agreement” option pops again the agreement that the user must read and accept before launching SoFi.

3.3.8. Credits
“Credits” option provides an e-mail address and a phone number for further support.
3.4. **SoFi Pro license activation step**

Pressing on any tab or button on the main SoFi panel will automatically connect to the FTP server and validate your SoFi Pro license. Make sure you are connected to the Internet. Otherwise, the validation process fails, and you will be asked to use SoFi standard, as long as the license cannot verified.

The license button turns green for successful license authentication and shows how many licenses of your SoFi Pro package are currently in use. On the left side, the number of days left for the SoFi Pro license is shown in green. Below in gray the panel reports the remaining time prior to the next server connection (Figure 5).

To log off from the server press on the license button (green button). This option is useful in case only a few licenses are purchased, and SoFi Pro should be sequentially used on several PCs. To log on just press any button or tab on the SoFi panel as described above.

SoFi requires the Internet for the license validation and will connect to the server every 24 hours, once being logged in. During these 24 hours SoFi Pro works perfectly offline. Hence, if SoFi Pro must be used offline, make sure you validate your license first while being connected to the Internet.

![SoFi license panel with one SoFi Pro license activated (green panel). The example shows 61 days left for the SoFi Pro license (green text on the left of the button) and SoFi will attempt the next connection to the server for the license validation in one hour (gray text below).](image)

**Figure 5**
3.5. Data hierarchy in IGOR

Once SoFi is launched the folder SoFi with all its subfolders are created under the root folder (see Figure 6). The data folders contain:

“External_data”
The raw PMF input.

“Raw_input”
External reference profiles together with the associated variable bases (must be in txt format, see 5.2.1 for more details)

“Raw_pr”
The time series of external data together with the associated time stamp (see 5.2.2 for more details).

“Reference_pr”
The reference profiles that contain the same variables present in the PMF input matrix.

“Reference_ts”
The external information that matches the time of the PMF input (default resolution), SoFi Pro: in addition, hourly, daily, weekly, monthly and yearly resolution are also evaluated.

Note: The user should not be concerned about the length (time series) / amount of variables (profiles) of the external data. There is a matching algorithm (see 5.2) that makes sure that the external information is harmonized to that of the PMF input.

“Eff_input”
The PMF input used for the current PMF run.

“Results”
Once the model results are imported by the user, they will be stored in the folder “root:SoFi:Results:XXX”, with “XXX” being the name of the user-defined HDF5 file.

“Preanalysis”
Data related to the pre-treatment part (tab 2 in the main SoFi panel).

“Variables”
All variables are necessary for the software. The user is not supposed to modify anything in here!!!

Figure 6 Data structure for SoFi.
4. **Import and define the input data**

SoFi is structured in a task-dependent panel system. The main panel is shown in Figure 7. This is the panel that controls all major tasks in SoFi, such as defining the PMF input, creating various PMF cases, launching the ME-2 solver, importing and visualizing the results. Subpanels should be opened and closed by clicking on the respective “open” buttons. Closing panels and/or graphs using X (top right) could cause unexpected behavior of some features.

![Main SoFi panel, import tab](image)

**Figure 7** Main SoFi panel, import tab. The “Import data” button allows automatic import of the PMF input and external time series and profiles. Under “Define PMF input” the user has to define the PMF input.

### 4.1. Import data

The first step is to import the data into the proper folder of the current IGOR experiments. This can be done manually or automatically. SoFi currently supports automatic data import for the PMF input from IGOR text files (.itx) and excel files (.xls and .xlsx), as well as external time series and external profiles from IGOR text files (.itx), excel files (.xls and .xlsx), comma-separated values files (.csv), data files (.dat) and text files (.txt). Using the button “Import data” opens the import panel where the details for the automatic import can be specified. For the manual import, please consult the subchapters for the locations where the data has to be copied to.

Please contact [support@datalystica.com](mailto:support@datalystica.com) if you experience problems with importing your data. We are always happy to expand the import function.
4.1.1. **PMF Import**

The PMF import needs to be stored under “root:SoFi:External_data:Raw_input”. The PMF input should contain at least the data and error matrix, as well as the time series and profile (variable) vector. The error matrix can also be calculated by SoFi if the user imports from an excel file (see details below).

By the drop-down menu on the top, the user can select the data file type from which SoFi should import the PMF input. For .itx files, the required matrices should be stored as separate 1D or 2D waves, so that it can be imported automatically and no further information is required. The “Import” button opens a file dialog to select the .itx file.

For PMF inputs from an excel file, the user first needs to select the file (via the “Choose file” button) and then specify the cells in which the data matrix, error matrix, time series and profiles can be found. In addition, the spreadsheets have to be specified. Data matrix and error matrix are allowed to be in separate spreadsheets. For excel files, SoFi imports the profile vectors based on the row names as a text wave and generates automatically a corresponding numeric wave, as needed to define the PMF wave. If the field for the error matrix is left blank, SoFi assumes that there is no error message and gives the option to let the error matrix (later named “Error_Mx_SoFi”) be calculated automatically. For this feature, SoFi supports three types of error matrix calculation:

“Polissar et al., 1998”

In the approach described by Polissar et al. (1998), data matrix and error matrix are composed as fixed fractions of the method detection limit (MDL):

For conc. > DL: conc. = measurements, $\sigma = \frac{1}{3} MDL$  \hspace{1cm} (14)

For conc. $\leq$ DL: conc. = $\frac{1}{2} MDL$, $\sigma = \frac{5}{6} MDL$  \hspace{1cm} (13)

“Norris et al., 2014”

In this approach uses an equation-based uncertainty and an error fraction for concentrations above the MDL. For concentrations below the MDL, the approach from Polissar et al. (1998) is followed. This error matrix calculation is also used in EPA PMF.

For conc. > DL: $\sigma = \sqrt{(error \ fraction \cdot conc.)^2 + (MDL)^2}$  \hspace{1cm} (16)

For conc. $\leq$ DL: $\sigma = \frac{5}{6} MDL$  \hspace{1cm} (15)

Both error matrix calculations require the input of the method detection limit (MDL) for each element/species/iron/etc. as illustrated in Figure 8. The equation-based uncertainty according to Norris et al. (2014), additionally requires the input of the error fraction. SoFi accepts an error fraction for each element or one error fraction for all elements. Both, input values and error fraction need to be separated by semicolons.
By checking the box “Average time series”, SoFi averages the imported PMF input to the desired time resolution (in minutes), storing both, the original input and the averaged input under “root:SoFi:External_data:Raw_input”. After the import of the original PMF input, a window pops up where the user has to define the desired time resolution (in minutes). For the import from an .itx file, the user has to additionally define the data matrix, error matrix and time series.

4.1.2. Time series import

The external time series need to be stored under “root:SoFi:External_data:Raw_ts”, which is done automatically if the “Import external ts” option is chosen. The minimum data that is required, is a time series and a corresponding numeric value series. In a first step, the user has to choose the file type from which the data should be imported. In the following the specifications for the different file types are explained:

- itx files: no further information is required, the “Import” button will open a file dialog to select the .itx file and all waves in this file will be imported.
in a first step, the user has to select the file by clicking on “Choose file”. The user needs to define where to find the time series (only one column at the time), by giving the start and end cell, e.g. A2,A40. Similar for the data (concentrations). The user has to again give the start and end cell, however here more than one column is supported, e.g. B2,H40. The spreadsheet name also has to be defined. SoFi will import the data column-wise.

Row-wise import is not supported.

the user has to choose the file by clicking on “Choose file”. First, the user has to define in which column the time series can be found. Please note, the column number starts from 0. In the following three options, the user should define the time format that is used. “Format” defines the date format, while “*” is the date separator (e.g. “.”, “/”, “-”) which has to be specified in “Date separator (*)” (without “*”). Whether the time series contains a time information and if so in which format, has to be selected under “Time”. In a next step, the columns which contain data need to be specified. “1st column of data” is the column from where the data should be imported (note, column numbers start from 0). “Nb of columns” defines how many columns should be imported. It can be either set to a certain number or to import all columns after the 1st data columns, set it to 0. To name the imported waves correctly, set the “Header (row)” on where to find the names (note, row numbers start from 0). If there is no header, Igor will name the concentrations waves “waveX”, with X being an increasing number. “Nb of rows to skip” defines how many rows should be skipped. The header is not automatically included. In “Column separator” the column separator needs to be defined. In case the columns are separated by tabs, please use \t. “Decimal character” is pre-defined as dot but can be changed into comma.

If data waves with the same name are being imported at different steps, they are not overwritten, but a continuous number is added to the end of newly imported wave.

4.1.3. Profile series import

The external profile series need to be stored under “root:SoFi:External_data:Raw_pr”, which is done automatically if the “Import external pr” option is chosen. The minimum data that is required, is a profile series (variable vector) and a corresponding intensity series. In a first step, the user has to choose the file type from which the data should be imported. The required specification for the import is similar to the time series import, except that there are not information about the time format required. Please consult the information under “time series import” on how to set up the import and replace time series by profile series. SoFi imports the
profile variable vector as a text wave, so that it can be directly used in the “treat external data” option.

4.2. Define PMF input

The internal string variables in SoFi must be linked to the just imported PMF data. This is done using the dropdown menus “Data matrix”, “Error matrix”, “Tseries nb”, “Variables nb”, “Variables txt” (Figure 7).

4.2.1. The data and error matrix

The error matrix must contain the word “_err” in the name. This allows for an internal differentiation between the data and error matrix. Large names, e.g. more than 15-20 characters should be avoided. This is a restriction given by the ME-2 solver.

4.2.2. The time wave

“Tseries nb” contains the numerical information of the time series.

4.2.3. The variable vectors

“Variables nb” contains the numerical information of the variables. “Variables txt” must be of text format and typically contains the text information of the variable, e.g. the name of the elements for elemental analysis or the name of the ions for HR-AMS analysis. However, even for unit mass resolution variables both vectors are required. In such a case the information is in both vectors the same but with a different representation, i.e. numerical for “Variables nb” and text information for “Variables txt”.

Generally suggested names are: **Mx_data, Mx_err, amus_nb, amus_txt, tseries_nb** for the data matrix, the error matrix, the numerical variable vector, the text variable vector and the numerical time series vector, respectively.

4.2.4. Data type

The dropdown menu allows specifying the details of the PMF input data, such as the type of data: ACSM / UMR-AMS, HR-AMS, filter data, others etc. This is important to define, if the CO2-related variables should be excluded (ACSM/AMS) (see 6.2.1) as well as to specify units of the PMF input. For ACSM and AMS the unit “µgm⁻³” is automatically passed. For the other cases the user is prompted to define the units manually. The unit information is used in SoFi any time the contribution is visualized.
4.2.4.1. **Minimum error for AMS data**

For the UMR-AMS and HR-AMS data the user can further apply the minimum error to the error matrix, if not already present. It is important to pass the wave with the exact name “minErr1ion” (generated from the previous data analysis software SQUIRREL or PIKA) under “root:SoFi:External_data:Raw_input:”. This is given, if an IGOR text wave containing the PMF input is imported using the “PMF input” button.
5. SoFi – pre-treatment

5.1. Treat missing data in PMF input

Every cell of the PMF data matrix represents an equation that is used during the ME-2 minimization process. Missing information in the PMF input must be treated accordingly and excluded from the set of equations. Two cases can be distinguished, a) entire row / column has no information, b) few missing data.

The two cases should be treated in this order a) than b) and the first step is to define the symbol for missing information under “Define symbol for missing data”.

a) Entire row / column has no information

Note that entire rows / columns are removed directly from the main imported input under “root:SoFi:External_data:Raw:input”, once the button “initialize PMF input & check for missing data” is pressed. The process is repetitive, in that if various symbols are present for missing information, e.g. NaN and 0 for AMS data, the symbol can be changed using the corresponding button and the initialization process can be repeated. This action also creates a copy of the current PMF input under “root:SoFi:Eff_input”.

If more symbols representing missing data are present, the order matters and the user must start with the one that is uninterruptedly present throughout an entire row / column. This allows to correctly remove entire missing rows / columns.

Figure 10 Main SoFi panel, pre-treatment tab.
b) Few missing data
The symbol used for tagging sparse missing information is supposed to be the last symbol present under “Define symbol for missing data” (Figure 10). For the tagging and the exclusion of these cells consult 6.2.3, III. Miscellaneous settings – missing data.

5.2. Treat external data

The external information is harmonized to match the PMF input data using the yellow panel by clicking on “Treat external data for the PMF run/result comparison” (Figure 10).

5.2.1. Profiles

The algorithm extracts the PMF variables present in the external vector(s) and stores the copy of the external vector(s) with the same length as the number of variables in the PMF matrix under “root:SoFi:External_data:Reference_pr”. Missing variables in the external vector(s) are defined as “NaN”s. The extraction works by comparing the text information, e.g. name of ions for HR-AMS of the PMF input with that of the external data. Therefore, the x-axis for the external data must be linked to the text information (Figure 11 left panel). Pressing the “SHIFT” key allows to select adjacent entries while the “CTRL” key allows the user to add non-adjacent entries.

FOR AMS/ACSM user

!!!When importing mass spectra for AMS or ACSM data, make sure that these externals match the fragtable of the PMF input. The extract procedure in SoFi will fail otherwise!!!

Figure 11    Left panel: extraction of reference profiles. Right panel: re-grid panel for collocated external time series.
5.2.2. Time series

The algorithm re-grids the time series of the external data. The following options are available:

“EXTRACT (same timep.)” In case the time stamp of the external data equals that of the PMF input with the presence or absence of certain points.

“INTERPOLATE” In case time resolution of the external wave is similar but not equal to that of the PMF input.

“AVERAGE” In case the time resolution of the external wave is higher, e.g. PMF input resolution 30 min and external resolution 5 min.

(start/middle/end sav. p.) defines where the data should be stored, either as the start / middle / end saving time.

SoFi Pro: The options hourly, daily, weekly, monthly and yearly resolution (as shown in Figure 12) are automatically evaluated. This option is very useful, when

- comparing PMF results with hourly, daily, weekly, monthly and yearly resolution
- quick inspections of PMF results, where the analysis with a high temporal resolution would require too much waiting time.
- comparing PMF results to external data that has a lower time resolution, e.g. PMF on AMS and 24 hours filter data as external data. Then the options hourly or daily can be used for comparison.

![Figure 12](image)

**Figure 12** Sanity plot of temperature over time for default resolution (black, instrument resolution), averaged using the resolution of the PMF input (red, raw resolution), hourly resolution (green) and daily resolution (blue). Various resolutions can be enabled/disabled from the control bar on top.

The algorithm stores a copy of the external wave(s), with the same number of points as in the PMF input matrix under “root:SoFi:External_data:Reference_ts:raw” as well as the corresponding hourly, daily, weekly, monthly and yearly copies under “root:....hourly”, “root:....daily, etc.”.
Missing variables are defined as “NaN”s. The x-axis is the regular time series information in numerical representation.

It is convenient to add the units of the external time series (Figure 11 right panel). This information is later used when the corresponding external time series is plotted in the result section. Pressing the “SHIFT” key allows to select adjacent entries while the “CTRL” key allows the user to add non-adjacent entries. After saving the external wave(s), a sanity plot pops and compares the raw time series and the re-gridded time series (Figure 12). This is performed for every chosen external time series.

5.2.2.1. Wind data
Wind data (wind direction and wind speed) are treated specially when averaged in SoFi. Wind direction is averaged by averaging the x and y components of the angles separately and merging them back together after having averaged (degree dependent averaging). For wind speed information the normal average as well as the degree dependent average as described above for the wind direction is applied.

Note: wind direction is only averaged degree dependent when the name of the wave with the wind direction information is “wind_dir”. Wind speed is averaged degree dependent when the name of the wave with the wind velocity information is “wind_vel”. Moreover, the degree dependent averaging for the wind velocity is not directly popped on the “Sanity_plot” graph (this graph shows the normally averaged wind velocity), but the wave is generated and stored under the corresponding Time series folder with the name “wind_vel_degdep”.

5.3. Explore PMF input

It is highly recommended to perform a variable exploration prior to the PMF analysis. The user should get a feeling, especially the key variables, how they vary over time. For this reason, an additional panel was developed that pops after clicking on “Explore trend of the variables” (Figure 10). The explore data panel (green panel, Figure 13) has two adjacent subpanels that govern the options on the graph and on the data. The following options are available:

5.3.1. Options for the variable(s)

“y-var.” a list of all variables of the PMF input matrix including external data
“x-var.” time, a list of all variables of the PMF input matrix including external data
“color-var.” normal (no wave), time, hours, any variable of the PMF input matrix including external data
“t-plot” normal time series, diurnal cycle, weekly cycle, monthly cycle, yearly cycle
“type” abs, fraction (a variable normalized by the sum of all input variables), signal to noise (S/N) (consult 3.3 for the S/N calculation)

“select based on class / time” see below for more information

Plotting the variables uses a string search algorithm in the background and therefore the names of the variables (in the variables text wave) must differ. If for example elements from various size fractions are present, then the element name must contain the information of the size fraction too, e.g. “Na” for the element in both size fractions is not sufficient, it should be “Na fine” and “Na coarse”, or a similar specification.

Figure 13 Panel for exploration of the input variables. Left part shows the graph and the subpanels on the right govern options of the graph and data.

Cycle plots in SoFi report the information in “start saving time”. The data of e.g. a daily cycle stored under 2 pm is representative for the data between 2 and 3 pm.

5.3.2. Select based on class / time

This selection panel allows the user the easily and efficiently inspect data based on a multitude of options summarized in Figure 14. Here the options are discussed based on the three tabs separately.

“time-selection” selection of hours, days, months and relevant years

“overview over time” user can select several regions of interest with the marquee (draw a marquee around the period of interest (only points with both x-
and y-coordinates falling in the marquee will be considered), click on the right mouse button and confirm under the menu “marquee”. Note: this option is additive, i.e. the user can select and confirm various regions several times with the marquee tool.

The user can select under “time series wave” also external waves. This allows selecting points not only based on the PMF input, but also on external waves, e.g. investigating variables during episodes with high concentration of a tracer.

All entries are selected as default, i.e. all points are used.

the user can specify points to belong to certain classes and select among these when present from this tab (see 5.4.1.1 for more details on the class)

![Image of the three tabs of the selection subpanel based on time / class.](image)

**Figure 14** The three tabs of the selection subpanel based on time / class.

### 5.3.3. Options for the graph representation

“Pop graph” pops the current representation from the panel.
“Close graph(s)...” closes all graphs related to the exploration panel
“type” user can choose between various default trace representations, e.g. solid line, dashed line, etc.
“size” size of the trace
“repr.” Option only available for cycle plots, where mean or median and the corresponding spreads (standard deviation, interquartile range (IQR), IQR with 10th and 90th %, IQR with percentiles and min/max) can be added to the representation
“use this color/define color” user can define color for the current trace. If nothing is specific, SoFi uses colors from the color palette starting with red and moving row-wise and using every 2nd color (avoids using similar colors twice).
“linear fit” performs a linear fit, makes only sense when performing a scatter plot
“error-x..” error also assumed for the x-wave and least squares are not vertical but rather orthogonal to the fit (ODR = 2 in IGOR)
“through zero” fit is forced to go through zero, i.e. intercept = 0
“New” / “Add” Chosen data can be plotted using the button “New”, or added to an existing graph using the button “Add”.
“Remove last” The last trace is removed from the graph

Two graphical examples for this exploration panel are shown in Figure 15.

Figure 15 Daily cycle of m/z 57 mass (left), scatter plot of f44 against f43 colored over time.
5.4. Prepare PMF input for the run

5.4.1. Select or modify effective PMF input

5.4.1.1. Class(es)

Variables or time points can be grouped together into classes. Classes are a powerful tool for:

“PMF input” PMF can be performed over certain classes of variables or time points only, e.g. PMF over air masses affected by continental air (back-trajectory data would be a class wave).

“C-value” classes can receive different relative errors when performing PMF (see 6.2.4.2 for more details). This option is only available for variable classes.

“Result inspection” PMF results over Profile / Time series can be shown for defined classes only

Classes can be defined in two different ways in SoFi.

Import from preexisting file

Classes have already been defined beforehand and need to be read and linked in SoFi. Class waves can exist for variables or time points and the user governs the option based on the “Profile, Time series” dropdown menu on the “PMF_input_sel” panel (upper red rectangle on Figure 16). The dropdown menu “class names (txt wave)” reports the list of waves contained under “root:SoFi:External_data:Raw_input:”, where the unprocessed PMF input is stored (lower red rectangle on Figure 16). The class waves must be in text format and possess the same length as the PMF input (over variables / time points) (see Figure 17 for an example on the variables). It is important that all variables / time points belong to a specified class, if a class wave is read. If not, the user must complete defining the untagged variables / time points directly in SoFi.
Class name wave for the variables is selected.

Example with 8 variables measured with different instruments (the same structure is valid for the time points)

<table>
<thead>
<tr>
<th>variable</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrate</td>
<td>Filter</td>
</tr>
<tr>
<td>Sulfate</td>
<td>Filter</td>
</tr>
<tr>
<td>Chloride</td>
<td>Filter</td>
</tr>
<tr>
<td>m/z 43</td>
<td>ACSM</td>
</tr>
<tr>
<td>m/z 44</td>
<td>ACSM</td>
</tr>
<tr>
<td>m/z 55</td>
<td>ACSM</td>
</tr>
<tr>
<td>m/z 57</td>
<td>ACSM</td>
</tr>
<tr>
<td>m/z 60</td>
<td>ACSM</td>
</tr>
</tbody>
</table>

8 variables (left column) and the corresponding class wave (right column).

For a PMF input containing these 8 variables the class wave has the same length and the respective entries “Filter” or “ACSM” in it.

Define classes directly in SoFi
If no classes are defined from external class waves, all variables / timepoints belong to the default class “all entries”. The user can define classes for variables in SoFi by entering into “type index” single or ranges of index for variables or timepoints.

Example

Making a new class for the variables at position 0 to 10 including the variables at position 12 and 15 requires the following entry: 0-10;12;15 (red rectangle on Figure 18)
Figure 18  defined class for the variables at position 0-10 including variables at position 12 and 15. This class received the class name “class_I” and a green color.

Clicking on “create and add a class” prompts the user to define the name and the color for the new class. Figure 18 shows the class for the discussed example. Currently, in the example the class “all entries” is still present, as only some variables have been classified so far. As soon as all variables or timepoints are members of user-defined classes, the default class “all entries” disappears. The user can inspect the grouped variables or timepoints under the tab “Visualize class(es)” (see Figure 19).

Figure 19  Inspection of the just defined class “class_I” as graph (left) and in table representation (right). The plot on the left colors the variables with the user-defined color. All other variables that currently still belong to the default class “all entries” are visualized in black.
Defining classes should occur before the blacklisting process as the class txt waves always refer to the entire PMF input. First blacklisting and then defining groups of variables or timepoints could cause some unexpected problems when visualizing the results later in the result panels.

5.4.1.2. Input data

Input data used for a PMF run can be modified using the blacklist function. This is advantageous, when points in time (spikes due to transient events, electronic noise or a PMF over only specific points is sought) or some variables (problematic and/or unnecessary variables, e.g. very noisy variables or variables with very little variability or a PMF over only specific variables is sought) shouldn’t be considered for a PMF run. The PMF input (blue panel) shows the current PMF input, whereas blacklisting is performed on the adjacent subpanel (see Figure 20). The selection subpanel contains the same options as those already presented for the subpanel under 5.3.2 and the user is referred to this section for the overview of the options.

Figure 20  PMF input panel useful for controlling the user-specified PMF input and for tagging points in time or variables into classes.

The default case considers all points in time and all variables for the PMF input. Removing some points or variables requires the user to select options either using checkboxes, the marquee function or the cursor (for variables only) on the subpanel. Once the selection is completed (it is an additive process and must be performed for the profile and the time series separately) (see Figure 21 for an example over time), the user can blacklist the selection by pressing on the button “BLACKLIST / REMOVE SELECTION...” on the third tab on the adjacent subpanel (see Figure 22). This will update the graphs and tables from the PMF input panel and the blacklisted data is supposed to disappear (see Figure 23).
Figure 21  (left) example where points between 0 and 2 am during the weekends for the months January and February for 2012 are blacklisted / removed. (right) example where two distinct episodes using the marquee tool are blacklisted / removed.

Figure 22  PMF input panel subpanel showing the third tab with the “BLACKLIST / REMOVE SELECTION...” and the “RESET INPUT..” buttons.

Figure 23  (left) default PMF input, i.e. all points over time and all variables are considered. (right) PMF input after being blacklisted by the user.
Blacklisting removes the selection, i.e. the user is supposed to select what should not be present later on in the PMF input.

If some points have been erroneously removed, the only way to restore the PMF input is by clicking on “RESET INPUT” from the third tab. This button calls the option “Initialize PMF input & check for missing data” (Figure 10) and reinitializes the entire PMF input matrix (see 5.1).

**SoFi Pro**: user-specific PMF input data can be stored and loaded from IGOR txt files directly from SoFi under “Input file” (see Figure 20). Saving the PMF input will also store user-defined values for $F$ or $G$, if already present and their corresponding $a$-values (see 6.3.1 and 6.3.3). This allows the user to store specific PMF input settings or initial values and constraints that can be loaded and tested, later on, if desired.
6. **SoFi – PMF run(s)**

6.1. **Define path, nb. of .exe and name of HDF5 file**

The “PMF run” tab is fully dedicated to set a specific PMF call using the PMF input and error matrices defined under the tabs “Import data” and “Pre-treatment” and to finally call ME-2 (see Figure 24).

6.1.1. **Path**

The user gives the location where the Me2.exe file resides by pressing the button “folder of executable”.

6.1.1. **nb. of executables**

The user defines the number of ME-2 instances that will run in parallel. This number must be less than or equal to the number of cores in your PC.

Example: If you chose 4 cores, SoFi will assign to any idle ME-2 instance out of four instances a new task during the solving process.

It could happen that some tasks are calculated faster (or slower) than the others. Therefore, the DOS windows are not necessarily all popped at the same time. Moreover, the run numbers in the log file / command history will not necessarily be in increasing order.

![Figure 24](image.png)  
**Figure 24** Main SoFi panel, PMF run tab.
6.1.2. **HDF5 file name**

The user defines a proper name for the PMF run under “HDF5 file name storing the current PMF run”. If this name already exists, SoFi prompts the user to try with another name.

!!! The “Delete file” button removes the specified HDF5 from the OS. Despite a last warning prompt, the user should pay attention not to delete important information. SoFi does not control hidden back-ups of the HDF5 files !!!

6.2. **Define general model options**

The orange button “Define settings pertinent to the current PMF run” (Figure 24) pops the panel governing all PMF settings (Figure 25).

6.2.1. **Define number of factors**

- **“starting factor”** defines the minimal number of factors (p) investigated by the model.
- **“ending factor”** defines the maximal number of factors (p) investigated by the model.
- **“nb. of PMF calls”** is the number of calls with the identical settings. SoFi will automatically adjust this value for fpeak analysis (see 6.3.2 for more details), for systematic a-value sensitivity analysis (see 6.3.3 for more details) and for C-value analysis (see 6.2.4.2)

Example: “starting factor” = 1 and “ending factor” = 5 the model will return solutions with 1, 2, 3, 4, and 5 factors.

6.2.2. **Convergence criteria**

The parameters controlling the goodness of fit during the PMF run can be configured directly from the table on the first tab (Figure 25). Model optimization is performed in three sequential stages; each stage corresponds to a row in the table. Final convergence is achieved only if the parameters of the third stage (last row) are fulfilled. These parameters can be modified by the user, and this is necessary to achieve convergence for e.g. large datasets, strongly non-zero fpeak values, etc.

There are two convergence criteria. The first criterion inspects the stability of Q between steps of the iteration process. If the difference in Q between the last and the actual iterative step is smaller than “deltaQ”, e.g. <0.1, for the number of consecutive steps provided in the column “cons_steps”, the algorithm converges. The other convergence criterion is “gg2_steps”, where the norm of the squared Jacobian matrix (matrix containing all partial derivatives of Q with respect to the model variables) is compared to the threshold value provided, e.g. of 0.1. The “max_steps”
column is simply the number of maximal iterative steps allowed at a given stage. If this maximal number of steps is reached during stages 1 or 2, the model moves on to the next stage and continues to search for a convergent solution. If this however occurs during the final stage, the model returns a non-convergent solution. Non-convergent solutions remain empty after having been imported and this is also reported in the log file. The user should easily recognize non-convergent solutions and can thereafter rerun the model using less restrictive convergence criteria, e.g. increasing by 10 or 100 the values of “deltaQ” and/or “gg2_steps” of the three stages (they should be increased/decreased together).

![Image of PMF settings](image)

**Figure 25** PMF settings, tab “General settings.

### 6.2.3. Missing data

Missing entries in the data matrix defined at the pre-treatment stage (see 5.1) must receive a proper value for the ME-2 solver.

At this stage the PMF input must contain only one symbol for missing data, e.g. “NaN”, “inf”, “0” or “-999”, but not more than one.

Clicking under “A.I) Define values for matrix cells containing no data” prompts the user to define the symbol for missing data together with an uncertainty value for its error cell (see Figure 26). Missing data that receives values that are smaller than the threshold value defined under A.II) are automatically excluded from the minimization process in ME-2. For safety reasons, their errors are also replaced by a very high value, e.g. 1000 times the highest error.
Figure 26 Prompting the value for missing data, the negative entry and its corresponding error value.

6.2.4. Error matrix (general)

6.2.4.1. Linear Error
In SoFi, the linear error expression (C3 expression) present in ME-2 can be added and enabled for the PMF calculation (Figure 27). The current error matrix will be updated accordingly:

$$error\ matrix[i][j] = error\ matrix[i][j] + C3 \cdot data\ matrix[i][j]$$  \hspace{1cm} (17)

Clicking on “A) Linear error (C3) based on ‘input data’” allows the user to define the C3 value. Once the C3 variable in IGOR is different than 0 (default), the C3 error button turns green and the C3 value is considered.

Figure 27 PMF settings, tab “General settings”, error matrix (general).
6.2.4.2. Relative scaling

Relative scaling is useful in case the analyst has reasons to believe that the error of certain variables needs to be increased / decreased compared to others. Several causes could legitimate this approach. Two cases where this might be relevant are briefly discussed.

“various relative errors” data from various instruments has different relative errors, e.g. one instrument 0.1% and the other 10% errors. PMF will fit the data from the first instrument more, as its Q contribution is much larger.

“different amount of var.” one instrument has 100 variables, the other 10. Variables from both instruments have a proper error estimation, e.g. 5% error. However, PMF will fit the data from the first instrument more, as its Q contribution is much larger. This becomes even more relevant, if co-variation between variables of the same instrument is high.

Rescaling the error of certain variables is achieved using the definition of classes in SoFi. The variables that should be rescaled must be part of one class (see 5.4.1.1). The following options are available:

“select” The error matrix is untouched

“User-specific C-values” The user defines C-values that should be tested on the adjacent table popping after having chosen this option (see Figure 28). The C-value is used to divide the current error, as shown in the following formula:

\[ \sigma_{\text{new}} = \sigma_{\text{old}} \cdot \frac{1}{C_{\text{value}}} \]  

(18)

where \( \sigma \) represents the error to be rescaled. Default C-value is one, i.e. the new error stays untouched.
“user-specific C-values” option with the corresponding C-values in the table on the right.

A C-value can be specified for every defined class. If for example three classes have been defined, the adjacent table will pop three columns to be filled. The number of entries also defines the number of PMF runs. In the current example (see Figure 28) four runs are made, where the errors of the first class are left untouched, whereas the errors of the second class are systematically divided first by 1, 10, 100 and finally 1000, i.e. the second class receives 1, 10, 100 and 1000 times more weight during these runs, respectively.

See also 7.3.3.2 (example of combined criterion) for monitoring C-value runs using the criteria panel.

For further details on the manual technique of the C-value the user is referred to the paper discussing the weighing strategy as C-value between AMS and gas-phase data (Slowik et al., 2010; Crippa et al., 2013).
6.2.1. **Error matrix (weight)**

Briefly, if the estimated measurement errors are uncertain, typically the case around detection limit and especially, if the systematic part of the error (background or blank) is not considered or poorly estimated, one could consider increasing the error / downweight the error. The weight strategy is based on the signal to noise ratio (see 3.3). Following options are available for weighing the errors (Figure 29):

**Model type**
- **“select”** the error matrix is untouched
- **“step function”** the two-step function proposed in Paatero and Hopke, (2003) is applied to the error matrix. The user defines the steps and their corresponding weights.
- **“1/S2N function”** the error is continuously downweighted with the penalty function 1 / S2N for S2N lower than 1 and untouched otherwise. This idea was first discussed in Visser et al., (2015).

**Data type**
- **“averaged”** weighing the data is based on the averaged S/N ratio
- **“cell-wise”** weighing the data is performed on each cell separately.

![PMF settings, tab “General settings”, error matrix (weight).](image)

The S/N plot as well as the affected errors can be inspected using the buttons “inspect signal to noise plot” and “check for weighted data”, respectively (Figure 30). The weight can be applied and removed with the buttons “apply weight factor” and “undo weight factor”, respectively.
Figure 30  Signal to noise (top) and weighted errors (bottom) for the averaged data on the left and cell-wise on the right.

Depending on whether the user is treating AMS / ACSM data and the m/z 44 related peaks are left in the data matrix, the option for downweighing those variables is enabled (Figure 31). The user has the option to apply or remove the applied weight using the buttons “weight CO₂ related var.” and “undo CO₂ weight”, respectively.

Figure 31  PMF settings, tab “General settings”, error matrix (weight), showing the CO₂-related buttons.

Many of the just mentioned free parameters have not been systematically tested and hence no recommendations can be given at this stage. There are plans for a study that will address the sensitivity of PMF results with respect to these settings. More details with possibly some recommendations will follow soon here.
6.2.1. Other various options

6.2.1.1. Robust model threshold (A)
PMF is run in the robust mode to safeguard the PMF run from drifting away due to outliers (see 2). The user can enable or disable the robust mode and define the threshold for the robust criterion in section A) (Figure 32).

6.2.1.2. Model normalization (B)
The user can select whether the sum of each factor profile is normalized to one during the iteration (default case, enabled) or factor profiles are left unnormalized (disabled).

6.2.1.3. Mass closure (C)
There is the option to fit the sum of all factor contributions to a certain mass, e.g. PM$_{1}$, PM$_{2.5}$, PM$_{10}$ that can be chosen with the pop-up menu. The mass wave must be present under “root:SoFi:External_data:Reference_ts:raw” (consult section 5.2.2). Selecting a possible time wave changes the red button “mass close disabled” to the green button “mass closure enabled”. This adds mass balance equations to $Q'_{aux}$ and those are simultaneously fitted with all other equations defined by the regular PMF algorithm $Q''$ (Figure 32).

This option should only be used, when the mass difference is very small. The user is not supposed to regularly fit “invisible” mass. It is also recommended to run the model without fitting the missing mass and to compare the mass differences.

6.2.1.4. Multi-time (D)
The multi-time feature offers the possibility to run the PMF algorithm on a dataset containing two and more classes of data that differ in temporal resolution (Ogulei et al., 2005). The goal is to utilize the native resolution of each of the classes of instruments without being forced to average beforehand but rather consider and taking care of the different resolutions when performing PMF. The PMF equation that is solved instead of (1) is the following:
\[ x_{sj} = \frac{1}{(t_{s2} - t_{s1} + 1)} \sum_{k=1}^{p} f_{jk} \sum_{i=t_{s1}}^{t_{s2}} g_{ik} \eta_j + e_{sj} \]

(19)

where \( x_{sj} \) represents a single point of the model matrix \( X \) for the instrument sample \( s \) and the variable \( j \), \( t_{s2} \) and \( t_{s1} \) are the start and end time of the sample \( s \), respectively, \( f_{jk} \) is the mass fraction of variable \( j \) from factor \( k \), \( g_{ik} \) is the mass contribution of factor \( k \) for a given time-point \( i \) for the variable \( j \), \( \eta_j \) is the adjustment factor for the variable \( j \) and \( e_{sj} \) is the non-modelled part of the residual matrix \( E \) for sample \( s \) and variable \( j \).

The sample with the highest temporal resolution represents the main equations in \( X \). For the main equations \( t_{s2} \) and \( t_{s1} \) are identical and the inner sum disappears, as well as the normalization fraction at the begin. For the other samples, the point \( x_{sj} \) is represented as the average over all points from the main equations present during the sampling period, i.e. between \( t_{s2} \) and \( t_{s1} \), as stated in equation (19). Adjustment factors \( (\eta_j) \) for the main equations are also modelled during the iteration and are typically between 0 and 1. \( \eta_j \) for the samples with lower time-resolution are set to unity, as they normally provide better sampling and analytical accuracy. In addition to solving equation (19), the Multilinear engine applies a smoothing equation to eliminate unrealistic residuals when fitting data from lower and higher time-resolution at once, based on the following equation:

\[ g_{i+1k} - g_{ik} = 0 + \varepsilon_i \]

(20)

where \( g_{ik} \) is the source contribution of factor \( k \) during the time-point \( i \). The user can vary the strength of this smoothing equation based on the smoothing value (default value is 0.01) next to the multi-time option (Figure 32). The idea behind is that when a source contains some high-resolution species, reducing the residual in equation (20) leads to an increase in the residuals of equation (19) and the high temporal variations tend to be conserved.

In case the measurement for the classes with lower time resolution contains some gaps, then it is extremely important to pass the information of the start and end time for these classes. In order to pass this information, the user MUST pass the PMF input time series as a 2D matrix, where the first column contains the start time and the second column the end time. Note that the start time for the main/default class is not used, hence for this class the time series input wave can (in case it is easier for the user) contain two times the same information.
In order to perform multi-time, the user is supposed to prepare and read the input and error matrices in SoFi as described with the following schematic:

![Diagram](image)

**Figure 33** Input data and error matrices for multi-time

where sample 1 has the highest time-resolution and sample 2, 3, 4, etc. have always lower time-resolution. The empty/white parts in the final matrix (right Figure) are all “nans”. The time series of all samples are past in their native time-resolutions. This means that the time series wave will move back and forth, as the samples 2, 3, 4, etc. are supposed to cover the same time-range as sample 1. SoFi will then find the proper index values and pass them to ME-2 for the additional averaging for the equations $x_{sj}$ for the samples 2, 3, 4, etc. given in the example in Figure 33. The points of the higher classes are expected to be within the main time series, i.e., that of the first class. If that is not the case, SoFi will automatically not consider these averaged points (gap-safe). Moreover, the multi-time code is also safe when used in combination with the bootstrap resampling technique (see 6.4.1).

The user is supposed to define the non-diagonal elements in the data and error matrix as missing values (please consult 6.2.3). In addition, the user must define a class for each instrument sample over time and over the variables (consult 5.4.1.1), the multi-time equations are not properly passed otherwise.

The resulting adjustment factors $\eta_j$ described before are reported as wave “adjustment factor” in the corresponding result folder: “root:SoFi:Results:HDF filename:Solutions:run_X”, for further manual inspection. Note that adjustment factors close to unity indicate a good agreement between the different measurements.
6.3. Define options for the model matrices

6.3.1. Seed

The seed option populates the model matrices G and F. Random initialization typically explores the amount of local minima, where the final Q-value is still high, compared to the global minimum or regions in the Q-space with almost identical minimal Q-value (rotational ambiguity, see 2.3). Increasing the number of PMF runs with completely random values for G or F augments the changes to find these regions.

![PMF options settings](image)

**Figure 34** PMF settings, tab “Model MATRICES”, seed.

The table in the tab “Model MATRICES” for the selection “seed” shows the overview of the chosen model entries for the matrix F “Profiles” or G “time series” (Figure 34). The default value is “r” for random. The user governs the various possibilities with the subpanel that appears after pressing on “Define model entries”.

6.3.1.1. Main settings

“Factor” all factors or a single factor position, e.g. factor 1 can be chosen

“Initial model values” can be one of the following options:
random (r): always completely random entries
pseudo random (pr): randomly initialized at the beginning of a PMF call, and the same random value is used through all other runs belonging to the PMF call
from vector (fv): solution matrix $F / G$ is populated using external information from the folder “root:SoFi:Reference_ts:raw or root:SoFi:Reference_pr”. Consult 5.2 for preparing external data directly in SoFi.

![Image](image.png)

**Figure 35** Panel for initial model values from vector. HOA_Paris profile normalized to 1 is going to be written on factor position 1 for the profile matrix $F$.

6.3.1.2. **Options for “from vector (fv)”**
When choosing “from vector (fv)” the panel “model_define” (Figure 35) 4 additional options are available. For each external vector one can either choose all variables of the vector or a specific
variable to populate the model matrix. The choses variable(s) of is (are) shown in the left table “usable variables from external vector”. The entries of current model column which are still undefined are displayed in right table “unspecified model variables”. The tables get updated each time some information is passed to the model matrix by clicking the button (write into model matrix”. Writing information to the model matrix is additive. If the vector of the matrix $F / G$ contains some undefined entries, the button “Some var. of factor still unspecified” remains yellow. As soon as all entries are defined, this button turns to green “All model var. of factor defined”.

The user has the option to normalize the entry to the value provided under “Sum factor var.”. This option is linked with the information that is written and not to the final vector present in the model matrix.

**Example**

The user wants to normalize the sum of a profile vector (matrix $F$) to one and the information of this vector is stored in two independent external vectors, $A$ and $B$. Then the following calculation should be commuted:

$$w \cdot (\text{sum}(A) + \text{sum}(B)) = 1$$

(21)

w representing the weight value.

$$\text{nor. value} (A) = w \cdot \text{sum}(A)$$

(22)

$$\text{nor. value} (B) = w \cdot \text{sum}(B)$$

(23)

with $\text{nor. value} (A)$ and $\text{nor. value} (B)$ being the values entered under the voice “Normalize factor var.” on the panel “model_define” (Figure 35).

The button “write into model matrix” writes the chosen settings and the user can monitor the changes on the main panel (Figure 34). The user can write either an entire external profile or selecting certain variables that should be written using the option “Variables”. Writing is additive, i.e. the user can populate for example a vector of the model matrix based on more than one external vector.

If an entry was erroneously defined, the button “remove last settings” removes the last writing. The button “reset entire matrix” resets the current model matrix to default, i.e. random (r) for the entire matrix (Figure 35).

**Writing** the information of one or more external vectors to the model matrix $G$ or $F$ from the “seed” option does not mean that this information is constrained. This information simply represents the starting condition for the PMF run(s). Constraining this information or data that is derived upon this information is performed using the other two techniques in: “a-value” or “pulling equations”, see 6.3.3 and 6.3.4 for more details.
6.3.2. fpeak

The following fpeak options are available:

“disabled”  
fpeak rotations are disabled.

“global fpeak”  
enables rotations that affect the entire model matrices.

“individual fpeak”  
enables rotations that affect only chosen contributions, profiles.

For these fpeak options the following types of analyses are supported:

“exact fpeak”  
extact fpeak value is used.

“sensitivity fpeak”  
fpeak values are systematically tested between an initial and final fpeak using a user-defined fpeak resolution.

“random fpeak”  
fpeak values are randomly tested between an initial and final fpeak with a user-defined fpeak resolution.

6.3.2.1. Global fpeak

The global and the individual fpeak are those described earlier under 2.4. The user must enter the initial and final fpeak and the fpeak step interval. The range of fpeak values should be selected in a way that allows to fully investigate the range of solutions that are environmentally interpretable and at the same time are of mathematically similar quality (i.e. similar Q value) (Figure 36).

Example

One might explore the range of fpeak that leads to a 10% increase of the normalized scaled residuals ($Q/Q_{exp}$). A number of past studies have simply investigated a predefined fpeak range, e.g. fpeak values between -1 and 1, regardless of its effect on Q. This approach should be avoided, since the fpeak rotation is a function of the model and therefore, $Q/Q_{exp}$ will vary based on the input data and the model result.
Figure 36 PMF settings, tab “Model MATRICES”, fpeak.

Selecting one of the two options, “global fpeak” or “individual fpeak”, enables the fpeak rotations for the PMF run(s).

6.3.2.2. Individual fpeak

The user defines the initial, final fpeak and the fpeak step interval as for the global fpeak. In the cells of the matrix (Figure 37) passing “NaN” tells SoFi to pass the current fpeak during the PMF run(s). Conceptually, it can be regarded as the scalar with which time series of one factor is added or subtracted (based on the sign of fpeak) from another factor time series. The notation follows the rules of linear algebra.

Example

If for example the entry $T[0,1] = -5$, this implies that the time series of factor 1 is subtracted by magnitude -5 from the time series 2. Vice versa the factor profile of factor 2 is added to the factor profile of factor 1 by the same magnitude with opposite sign, 5.

However, a zero entry in $T$ does not prevent addition/subtraction from occurring, but rather indicates that this rotation, should it occur, is neither favored nor penalized by the model.
Figure 37  PMF settings, tab “Model MATRICES”, individual fpeak enabled with the example for T[0,1] as “NaN” for the fpeak value.

6.3.3.  $a$-value

The $a$-value option allows constraining one and more entries of the profile matrix $F$ or the time series matrix $G$ (see 2.4.4). The constrained information can be monitored on the table / graphs of the panel “PMF_options” (Figure 38). Adding or changing the constraints is done using the adjacent subpanel “constraint define” (Figure 39).

To enable the $a$-value constraints, use the dropdown menu below the table. The $a$-values are only used by the ME-2 solver when the button “A-value constraints are enabled” turns green (Figure 37).
PMF settings, tab “Model MATRICES”, α-value.

6.3.3.1. Main settings

“Factor” all factors or a single factor position, e.g. factor 1
“Type of constraint”
   “exact α-value”
   “1st dim. sensitivity”
   “2nd - 5th dim. sensitivity”
   “random”
   “pass limits”
“Variables” model variables, external variables, single variables, e.g. m/z 44

Some more details on the options are following.

6.3.3.2. Exact α-value (exact)
An exact α-value, e.g. 0.1 is used to constrain the defined information (profile / time series)

6.3.3.1. 1st dim. sensitivity (1dim. sens)
The α-value is systematically changed from “initial α-value”, using the step interval given by “α-delta” up to “final α-value”. This allows studying the PMF outcome as a function of the α-value in one dimension.
If you erroneously define two different settings for sensitivity analysis in one dimension, SoFi will only use the settings that are encountered first.

Example
First factor has a final $a$-value of 1 and second factor a final $a$-value of 0.5 and both factors have the constraint: one dimensional sensitivity analysis.
For the one-dimensional sensitivity analysis the final $a$-value passed to ME-2 will be 1, since the first factor is internally evaluated before the second one.

6.3.3.2. 2nd dim. sensitivity (2dim. sens) to 5th dim. sensitivity (5dim. sens)
The $a$-value is systematically changed from “initial a-value”, using the step interval given by “a-delta” up to “final a-value”. This allows studying the PMF outcome as a function of the $a$-value in a 2nd up to 5th dimension.

If only a sensitivity analysis in one dimension should be performed, the first dimension must be chosen. Using higher dimensions and keeping previous dimensions empty or non-defined generates an error message in SoFi.

Example
The first factor has a final $a$-value of 1 and a $a$-delta of 0.2 and performs a one-dimensional sensitivity analysis. The second factor has a final $a$-value of 0.5 and a $a$-delta of 0.1 and is linked with the two-dimensional sensitivity analysis. SoFi will perform \( \frac{1}{0.2} + 1 \cdot \left( \frac{0.5}{0.1} + 1 \right) = 36 \text{ runs} \)

6.3.3.3. Random $a$-value (random)
Random $a$-values are passed from a uniform distribution between zero and “Final a-value” with a precision given by “a-delta”.

6.3.3.4. Pass limits directly to ME-2 (limits)
If for a specific source, e.g. a traffic source the range of the profile is known, this range (limits) could be directly passed to ME-2 using this option. Two vectors representing the lower and the higher limits of the profile / time series are provided and ME-2 optimizes the profile during the PMF run within these limits. Note that a seed vector must be defined too. Use the lower or higher vector as starting condition when defining the corresponding seed information.

6.3.3.5. Constraining variables
The user has the option to choose which base of variables is used to pass the information of the constraint. Constraining variables is additive, as for defining the variables (see 6.3.1).

6.3.3.6. All model variables
This option uses as base the variables defined in the model matrix $F / G$. This leads to the same $a$-value constraint throughout all variables.
6.3.3.7. All model variables of external vector

This option uses as base the variables of an external vector. This is useful, when various \( \alpha \)-values should be used within a constrained vector (coming from various external profiles), e.g. if some variables of the constrained vector are more accurate than others. In such a case the less accurate ones could receive a higher \( \alpha \)-value.

Similar to the “model_define” panel (Figure 35) the “constraint_define” shows the variables that can be constrained (left table “variables that can be constrained”) as well as the ones that are still unconstrained (right table “still unconstrained variables). The latter table gets updated each time some information is passed to the constraint matrix. If the vector of the matrix \( F / G \) contains some unconstrained entries, the button “Some var. of factor still unconstrained” is yellow. As soon as all entries are constrained, this buttons turns to green “All model var. of factor constrained”.

Figure 39  The HOA_Paris profile defined on factor 1 of the matrix \( F \) earlier (see Figure 35) is going to be constrained with a fixed \( \alpha \)-value of 0.1.

The button “write into model matrix” writes the chosen settings and the user can monitor the changes on the main panel (Figure 38). If an entry was erroneously defined, the button “remove last settings” can be used to remove the last writing. The button “reset entire matrix” resets the current constraint matrix to default, i.e. empty entries (“-”) for the entire matrix (Figure 39). Every time the user wants to “write into model matrix”, SoFi checks, whether undefined information is going to be constrained and constraints only existing information.
It is recommended to constrain an entire profile / time series. If a wave is only partially constrained, its unconstrained part could grow during the PMF run and become dominant upon normalization in SoFi.

6.3.4. Pulling

The pulling option allows constraining one and more entries or combinations of entries of the profile matrix $F$ or the time series matrix $G$ (see 2.4.5). The constrained information can be monitored on the table / graphs of the panel “PMF_options“ (Figure 40). Adding or changing the pulling equations is done using the adjacent subpanel “pulling define” (Figure 41).

To enable the pulling equations, use the dropdown menu below the table. The pulling equations are only used by the ME-2 solver when the button “Pulling equations are enabled” turns green (Figure 40).

![Figure 40](image-url)  PMF settings, tab “model MATRICES”, pulling equations.
6.3.4.1. Main settings

“Factor”
a single factor position, e.g. factor 1
“Variables”
single variables, e.g. m/z 44
“Pulling type”
“exact pulling value”
“1st dim. sensitivity”
“2nd - 5th dim. sensitivity”
“random”
“pulling weight”
“strength with which the current value is pulled towards the user-defined anchor value”
“pulling Q”
“maximal allowed increase of Q”
“pulling anchor”
“user-defined anchor value”

Some more details on the options are following.

6.3.4.2. Exact pulling (exact)
An exact anchor value, e.g. 0.1 is used to pull the current information (profile / time series) during the iteration towards this anchor value.

6.3.4.3. 1st dim. sensitivity (1D)
The user-defined information (anchor or weight or Q) is systematically changed from an initial value, using user-defined step intervals to a final value. This allows studying the PMF outcome as a function of a pulling information in one dimension.

If you erroneously define two different settings for sensitivity analysis in one dimension, SoFi will only use the settings that are encountered first.

Example
First equation has a final anchor value of 0.3 and second equation a final anchor value of 0.2 and both equations have the pulling type: one dimensional sensitivity analysis.
For the one-dimensional sensitivity analysis the final anchor value passed to ME-2 will be 0.3, since the first equation is internally evaluated before the second one.

6.3.4.4. 2nd dim. sensitivity (2D) to 5th dim. sensitivity (5D)
The user-defined information (anchor or weight or Q) is systematically changed from an initial value, using user-defined step intervals to a final value. This allows studying the PMF outcome as a function of a pulling information in a 2nd up to 5th dimension.

If only a sensitivity analysis in one dimension should be performed, the first dimension must be chosen. Using higher dimensions and keeping previous dimensions empty or non-defined generates an error message in SoFi.
Example
The first equation has a final anchor value of 1 and a $\sigma$-delta of 0.2 and performs a one-dimensional sensitivity analysis. The second equation has a final Q value of 10 and a Q-delta of 2 and is linked with the two-dimensional sensitivity analysis. SoFi will perform $\left(\frac{1}{0.2} + 1\right) \cdot \left(\frac{10}{2} + 1\right) = 36$ runs

6.3.4.5. Random pulling values (RND)
Random pulling values are passed from a uniform distribution between initial value and final value with resolution of delta-value.

Figure 41 Variable m/z 58 of factor 1 undergoes a sensitivity analysis between 0.1 and 0.5 in interval of 0.1, with a pulling weight of 0.1 and a maximum allowed increase of Q of 10 units.

It follows the description of the pulling syntax under “pulling ANCHOR” that can be directly edited by the user.

Pulling SYNTAX
The left-hand side of the equation is linked with the model entries, either factor profile matrix $F$ or time series matrix $G$. The right-hand side defines the properties of the pull, i.e., the anchor, the strength and the maximum allowed increase of Q for this pulling equation.

$$\text{model variables to be pulled} = [\text{pulling value}];[\text{pulling strength}];[\text{pulling increase in Q}]$$

LEFT-HAND SIDE: model variables to be pulled
The user can define any mathematical expression representing a combination of cells of $F$ or $G$ that would also work form the IGOR command line.
some examples

1) pull of m/z 44 in factor profile 3
   factor_pr[44][3]

2) pull of ratio between m/z 44 and m/z 43 in factor profile 2
   factor_pr[44][2] / factor_pr[43][2]

3) pull of sum of m/z 44 in factor profile 2 and 3
   factor_pr[44][2] + factor_pr[44][3]

4) pull of sum between time points 50 and 51 in time series of factor 4

5) pull of sum of m/z 44 in factor profile 2 and 3 for the time point 100
   factor_pr[44][2] * factor_ts[100][2] +
   factor_pr[44][3] * factor_ts[100][3]

RIGHT-HAND SIDE: Pulling proprieties (anchor, weight and Q)
Three square brackets on the right-hand side of the equation are dedicated to the pulling details
of the anchor, the weight and the increase of Q. The full notation for each of these squares is the
following:

[“type”, “initial value”, “final value”, “resolution”]

The first entry of the square can be of type exact (type: “EXT”), sensitivity analysis (type: “1D”,
“2D”, “3D”, “4D” or “5D”) or random (type: “RND”). For “EXT”, there is no need for “final value”
and “resolution” and only “initial value” that is at the same time the exact value is passed, e.g.,
[EXT, 0.3]. For all other options all settings must be passed, e.g., [RND, 0,1,0.1].

In case something is wrong in the “pulling ANCHOR” line, one can use the “wipe away” button to
reset the entire line. The button “write into model matrix” writes the chosen settings and the user
can monitor the changes on the main panel (Figure 40). If an entry was erroneously defined, the
button “remove last settings” can be used to remove the last writing. The button “reset entire
matrix” resets all pulling equations.

These three parameters, i.e., anchor, weight and Q are treated as independent parameters.
Therefore, the user can perform independent sensitivity analysis for these three parameters, if
desired. This results in three independent sensitivity analysis dimensions, i.e., one is 1D, the
second is 2D and the third one is 3D.

Unfortunately, due to the syntax of the mathematical notation in ME2 not all the above-
presented pulling equations are working as is. In ME2, an equation accepts only the form:

Sum of terms, where terms are either single terms or products

-63-
Hence, for pulling equations describing ratios or for a term with two and more entries, the notation are slightly modified.

For fractions, the equation MUST be rewritten in a way to have the ANCHOR being equal to ZERO. This unfortunately, doesn’t allow to perform sensitivity tests on the ANCHOR for this type of pulling equation for the moment.

For a term containing a product, the multiplication symbol MUST be replaced by a COMA.

The following two examples illustrate how the pulling equation MUST be written for these two cases.

<table>
<thead>
<tr>
<th>Case</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>2) pull of ratio between m/z 44 and m/z 43 in factor profile 2</td>
<td>factor_pr[44][2] - 2*factor_pr[43][2] AND the ANCHOR must be set to ZERO</td>
</tr>
<tr>
<td>5) pull of sum of m/z 44 in factor profile 2 and 3 for the time point 100</td>
<td>factor_pr[44][2] , factor_ts[100][2] + factor_pr[44][3] , factor_ts[100][3]</td>
</tr>
</tbody>
</table>
6.3.1. **Factor specifications**

Each factor receives the default name “factor pos”. The user can change these names under “factor specs” (Figure 42) and give proper source names instead. When averaging PMF results in SoFi (see 7.3.2.1), factor possessing the same name will be averaged together. This allows the user to import various HDF files with various sources. The averaged result will then be based on all factor names defined by the user. As a consequence, the final factor specs list of the averaged solution might be larger than the factor specs list of a single HDF file.

**Example**

<table>
<thead>
<tr>
<th>HDF 1</th>
<th>HDF2</th>
<th>HDF3</th>
<th>avg. solution in SoFi</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOA</td>
<td>HOA</td>
<td>HOA</td>
<td>HOA form all files</td>
</tr>
<tr>
<td>COA</td>
<td>COA</td>
<td>COA</td>
<td>COA from all files</td>
</tr>
<tr>
<td>BBOA</td>
<td>SV-OOA</td>
<td>BBOA</td>
<td>BBOA from HDF1 and 2</td>
</tr>
<tr>
<td>OOA</td>
<td>LV-OOA</td>
<td>SV-OOA</td>
<td>OOA from HDF 1</td>
</tr>
<tr>
<td></td>
<td>LV-OOA</td>
<td></td>
<td>LV-OOA from HDF 2 and 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LV-OOA from HDF2 and 3</td>
</tr>
</tbody>
</table>

If an HDF file contains multi-factor solutions and the various factor solutions at the same position shouldn’t necessary be averaged together, then the user has the possibility to pass various names for a factor position.

**Example**

An HDF file with 4 and 5 factor PMF runs, where the 4th position in 4-factor runs represents something different then the 4th position in 5-factor runs. When averaging over 4- and 5-factor solutions, SoFi will use dedicated factor positions for OOA and LV-OOA. This leads to a six-factor solution for the averaged solution, since OOA and SV-OOA are separated from each other.

<table>
<thead>
<tr>
<th>HDF</th>
<th>avg. solution in SoFi</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOA</td>
<td>HOA</td>
</tr>
<tr>
<td>COA</td>
<td>COA</td>
</tr>
<tr>
<td>BBOA</td>
<td>BBOA</td>
</tr>
<tr>
<td>OOA;SV-OOA</td>
<td>OOA</td>
</tr>
<tr>
<td>LV-OOA</td>
<td>SV-OOA</td>
</tr>
<tr>
<td></td>
<td>LV-OOA</td>
</tr>
</tbody>
</table>

The user can also change the factor names in the HDF files before averaging the solutions. The factor specs list can be found here:
Consideration on the priority of the rotational tools

The explorative techniques for the rotational ambiguity are additive. The fpeak rotations can be enabled besides other techniques, such as the a-value or pulling equations or C-value approach, etc. The fpeak rotations have least priority, i.e. rotations are not allowed, if the resulting values are for example outside the limits given by the a-value constraints.
6.4. **Estimation of the statistical uncertainty**

The final PMF result / solution **must** also contain a quantitative measure of the statistical uncertainty. For ambient aerosol data the highest contributors to the statistical uncertainty are mainly random changes of the emission sources. Past studies ((Paatero et al., 2014) and (Ulbrich et al., 2009)) typically employed the resampling strategy “bootstrap” (Davison and Hinkley, 1997) to address and quantify this type of error for AMS data.

6.4.1. **Resampling strategy „bootstrap“**

The bootstrap method generates a set of new input matrices for analysis from random resampling of the original input data (Efron, 1979). This resampling perturbs the input data by including replicates of some points while excluding others (see Figure 43). Given that a sufficient number of resamples has been performed, the variation within the identified factors across all bootstrapped runs allows to estimate the statistical uncertainty.

![Figure 43](image_url)  
**Figure 43** Schematic of the default PMF input matrix (left) and a resampled one (right side). The colors represent the resampled rows. Note that the dark green color appears twice, whereas the olive-green present in the default PMF matrix is absent.

A crucial point when applying the bootstrap strategy on the PMF model is the presence of sufficient variability for retrieving the expected sources. If for example bootstrap is applied on high time resolved (order of minutes) PMF input data, it might happen that for an insufficient amount of bootstrap runs the points reflecting the rush hour are eventually missing and little variability is available to separate the traffic source. Hence, when performing only a few bootstrap runs (e.g. 50-100) one could make also use of the blocked bootstrap strategy (Norris, 2007), where blocks of data, e.g. entire days, are resampled together. In such a case, the timescale of the blocked bootstrapped data should be in the same order with that of the variability relevant for factor identification.
**Figure 44** PMF settings, tab “RESAMPLING strategy”

**Main Settings**

“unblocked” all points are treated separately when bootstrapping

“daily blocked” all points belonging to the same day are grouped (blocked) together.

More options will be added in future and the user will also be able to pass user-specific blocks (time wave with blocked data not necessarily in adjacent order).
6.5. Rolling mechanism

The main limitation of PMF is the modeling of static factor profiles that stand in contradiction with a dynamic system, where source profiles are supposed to vary over time. To account for a possible temporal variation of the factor profiles a rolling approach was suggested (Parworth et al. (2015), Canonaco et al. in prep.) for ambient AMS/ACSM data. The strategy involves spanning a PMF window and moving/rolling it over the entire dataset (see Figure 45). The PMF window must be smaller compared to the entire dataset, e.g. several days if the total PMF input comprises 1-2 years. More importantly, the window length should be chosen as such that the assumption of static factor profiles is legitimate.

![Schematic of a rolling PMF window.](image)

**Figure 45** Schematic of a rolling PMF window.

**Main Settings**

- “window width” size of the PMF window expressed in days
- “window shift” shift of the PMF window expressed in days
- “nb. of repeats” estimate of the nb. of repeats due to the overlap of windows

![PMF settings, tab “ROLLING mechanism”](image)

**Figure 46** PMF settings, tab “ROLLING mechanism”
6.6. Call ME-2 and perform PMF

SoFi calls ME-2 and passes the PMF input with the current instruction file together with the defined and constrained information, if present, to ME-2. The progress can be monitored with the Progress bar “ME-2 call” at the end of the main panel (Figure 24). The progress bar “SoFi” reacts on calculations performed in IGOR for SoFi only.
7. **SoFi – PMF result(s)**

7.1. **Define path of ME2.exe file**

The user gives the location where the ME2.exe file resides by pressing the button “folder of executable” (Figure 47). The “PMF result” tab is completely independent from the previous tabs and PMF results can be consulted even if they have been generated from another IGOR experiment or from another PC. The only restriction is given by the existence and the structure of the ME-2 folder under the OS. This is typically given, if the PMF run has been performed on the same PC. If the user wants to inspect some PMF results of another PC, then the HDF5 file containing the PMF results must be stored in the folder:

“ME2_Results” of the “ME2_engine” folder: \ME2\ME2_Results.

![Main SoFi panel, PMF result tab.](image)

7.2. **HDF5 file(s)**

7.2.1. **Select HDF5 file(s) containing the PMF run(s)**

The current SoFi version supports analysis of multiple PMF runs of a single PMF call but also among various PMF calls. This is accomplished by choosing all PMF calls that should be considered for the PMF analysis.
The button “II.a Select HDF5 file(s) containing the PMF result(s)” opens a subpanel where the user decides which of the HDF5 files are considered (Figure 48). Multiple PMF runs can be chosen. Pressing the “SHIFT” key selects adjacent PMF runs. Pressing the “CTRL” key allows the user to add non-adjacent PMF runs. Unselected PMF runs are not present in IGOR or if previously present, they will be removed from IGOR after having asked the user with a prompt message. This is necessary to keep the IGOR experiment small to avoid possible “out of memory” errors in IGOR. Removing the PMF runs from IGOR will obviously not affect their storage place on the HDF file(s).

![Subpanel showing the list of HDF5 files to be chosen for the PMF analysis.](image)

**Figure 48** Subpanel showing the list of HDF5 files to be chosen for the PMF analysis.

### 7.2.2. Import the chosen HDF5 file(s) into IGOR

Pressing on the button “II.b Import the chosen HDF5 file(s) into IGOR” will import the main information of the HDF5 file(s) to the corresponding folders.

### 7.3. Select and analyze PMF run(s)

#### 7.3.1. Select PMF run(s) from HDF5 file(s)

On the blue panel that pops after pressing “III.a Select PMF run(s) from HDF5 file(s)” (Figure 47), the user browses through the HDF5 files and selects the run(s) that should be kept for the subsequent PMF analysis (Figure 49).

7.3.1.1. **Settings**

- **“Choice base on”**
  - Manual choice (default case), automated selection (see below)

- **“Select HDF5 file”**
  - Browses through the list of the imported HDF5 files to choose the PMF runs

- **“x-axis type”**
  - User defines the x-label for the graph (see Figure 49) from the list: run nb., nb. of factors, fpeak value, a-value, pulling value, C-value
“avg. PMF run(s)” The PMF runs selected from the current or any other HDF file are averaged. Default is “no average”. This option is useful for the finalization. Details are discussed below in the section “criteria-based selection”.

7.3.2. Manual selection of PMF run(s) form a HDF file

For selecting a PMF run the user drags a marquee around the bar and confirms the selection by clicking on “Apply marquee in SoFi” in the context menu of the right mouse button. The choice “Reset marquee current” and “Reset marquee all” resets the option for the current HDF file or for all HDF files, respectively. The marquee selection is additive and can be performed on any of the graphs present on the panel “Select_PMFrms” as well as on the popped graphs. The graphs are the following: $Q/Q_{exp}$, absolute values, relative values, explained variation plot.

![Graph of PMF runs selection](image)

**Figure 49** Selecting a set of solutions from the HDF5 file “sensitivity_II” with the marquee.

Only on the $Q/Q_{exp}$ graph the color gray (default, unselected) turns to blue (selected). The gray bars (only x-value) representing PMF runs that should be further investigated must be in the marquee!
7.3.3. **Criteria-based selection**

It becomes difficult to fully explore a PMF run, in case hundreds to several thousands of runs are performed, e.g. when bootstrapping (see 6.4) or in a rolling window approach (see 6.5). In addition to a manual exploration of a PMF run (SoFi standard), SoFi Pro offers the possibility to reduce the complexity of a PMF run to a few proxies by defining so-called criteria. Hence, rather than inspecting the full PMF run the user visually inspects the scores of the defined criteria. Currently these criteria are defined as single points for each PMF run, e.g. a correlation coefficient, the ratio between two variables, the sum of some variables. This leads to one-dimensional plots of scores over all PMF runs. In future, more sophisticated scores and visualizations, e.g. storing an array or even a scatter plot for each PMF run will be provided (see below for more details). PMF runs with interesting scores/features can then be further investigated. To access the criteria-based selection the user must select the option “automated selection” in the drop-down menu “Choice based on:” (Figure 50).

![Select PMFruns panel showing the criteria-based selection button.](image)

**Figure 50** Select_PMFRuns panel showing the criteria-based selection button.

Clicking on the button “Criteria-based selection” launches the subpanel “PMF_criteria” (see Figure 51).
7.3.3.1. Type of criterion
The first tab in the “PMF_criteria” governs the definition and overview of the user-defined criteria (see Figure 52).

“point”
- average: single point is monitored for each PMF run
- sum:
- R Pearson
- R Spearman
- Kendall tau
- R uncentr

The various types of points are color-coded to allow the user to easily distinguish between them in the overview list “list of criteria”.

“1D wave” under construction: array is monitored for each PMF run. Future exploration goes through image plots

“scatter” under construction: scatter plot is monitored for each PMF run. Future exploration goes through movie plots
7.3.3.2. Options to set up a criterion

The subpanel "DATA"

The subpanel "DATA" (see Figure 53) contains the options for defining specific criteria. The following list is shown for a temporal criterion (ts). Options for criterion on profiles are characterized by (pr). Each type of data possesses a specific prefix or suffix (visualized bold in brackets below) that is internally used to differentiate between them. The user MUST NOT change these prefixes and in case entries are manually written to the criterion line these prefixes / suffixes must be written too. Otherwise, SoFi will not properly evaluate the criteria.

- "Time series, Profiles" to define the type of data (contributions (ts) or profiles (pr))
- "default, hourly, daily res." to define the temporal resolution
- "y (x)-type" solution (ts, pr), variable of solution (ts, …[var.]), external (ts, pr, ext_), variable (ts, var_), res. (ts, RES_), abs. res. (ts, ABSRES_), scaled res. (ts, RESW_), abs. scaled res. (ts, ABSRESW_), Q res. (ts, QRES_), explained variation (ts, explvar_), hours (ts, …[hour])
- "y (x)-var" varies depending on the "y-type" selection
- "normal, cycles" to define whether normal, daily, weekly, monthly or yearly cycles should be used
- "abs, fraction" to define whether absolute data or fraction, i.e. data normalized to the sum should be used

Figure 52  “PMF_criteria” with its subpanel “DATA”
The subpanel also gives the option to evaluate criteria based on class or time. Note, the computational time of the criteria will increase when a specific class or time has been chosen. Similar subpanels were already presented in 5.3.1 for the input data, 5.3.2 for time-dependent options. The reader is referred to those sections to learn more about these subpanels.

![Figure 53](image)

The subpanel “DATA” loaded with the settings (left) and time-dependent options for the factor time series (right).

The criteria list
Several different criteria can easily be defined in SoFi and the risk of losing track of which criterion was defined for which factor is high. Therefore, a small overview table showing the defined criteria is always visible from the first tab. The **color-code** of the type of criterion is also passed to the table. A batch table containing all information to a criterion, like exact time point chosen, type of data (abs or fraction) or resolution, can be consulted from the second tab (see Figure 64).
Several criteria visualized in the criteria list with the proper color-code. Criteria in bold are those that contain a temporal sub-selection. Highlighted criteria are used for repositioning of unconstrained factors.

Criteria that must be evaluated are checked. If a criterion has already been evaluated in a previous round, there is no need to re-evaluate it again. This saves a lot of computational time and makes the criterion analysis fast. Criteria that contain a sub-selection from the panel “Select based on class / time”, are visualized in **bold** in the criteria list. Moreover, criteria that should be used for repositioning (sorting) of unconstrained factors must be selected and are highlighted in the criteria list (see below under “The criteria approach on unconstrained factors” and “active versus passive criteria” for more details (see Figure 54).

**The criterion line and its SYNTAX**
The subpanel “DATA” can be used to easily add data to the criterion line that should be used for defining a criterion. Note that the features from this subpanel are typically additive. This means that existing information in the criterion line is not overwritten but new information is rather appended to the preexisting one.

In addition to this subpanel, the user can directly change, add or remove terms from the criterion line and perform simple mathematical operations that are also typically supported from the IGOR command line.

**Mathematical operations currently supported from the criterion line**

<table>
<thead>
<tr>
<th>mathematical operation</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>addition, subtraction, multiplication, division</td>
<td>wave1’+ wave2’, wave1’-wave2’, wave1’/wave2’</td>
</tr>
<tr>
<td>exponentiation</td>
<td>exp(wave1’), wave1’^2, 10^wave1’</td>
</tr>
<tr>
<td>square root</td>
<td>sqrt(wave1’)</td>
</tr>
<tr>
<td>min, max</td>
<td>min(wave1’, wave2’), max(wave1’, wave2’)</td>
</tr>
<tr>
<td>histogram(x min, x max, nb. of bins, wave)</td>
<td>histogram(-5, 5, 100, wave1’)</td>
</tr>
<tr>
<td>rules with brackets</td>
<td>(wave1’+wave2’)*(wave3’+wave4’)/wave5’</td>
</tr>
<tr>
<td>constant values</td>
<td>wave1’+wave2’+5’, wave1’+wave2’+5</td>
</tr>
<tr>
<td>Some combined examples</td>
<td>(wave1’+wave2’)/2+sqrt(wave3’^2)</td>
</tr>
<tr>
<td></td>
<td>min(histogram(-5,5,100, wave1’), histogram(-5,-5,100, wave1’))</td>
</tr>
</tbody>
</table>

Some more comments
In these examples wave1, wave2, wave3, wave4 and wave5 can be anything from the options presented under presented under “The subpanel DATA”.
For the expressions min(wave1, wav2) and max(wave1, wave2) the comparison is performed on a point-by-point basis for the two waves. The result wave has the dimension of wave1 or wave2 (they must have the same dimension) and contains for each point \( i \) the smallest value, i.e. \( \min(wave1[i], wave2[i]) \) is performed for all points \( i \).

This function diagrams(x min, x max, nb. of bins, wave1) creates the histogram for the specified wave1 with the parameters, x min, x max and nb. of bins. Selecting the option “fraction” normalizes the sum of the histogram to 1. Note that the histogram option can be applied only once per criterion. Moreover, histogram is interesting when for example comparing the residuals of the classes on a bin-by-bin basis in combination with the option min(), see below.

There are only few rules to the criterion line:

1) The pre and suffixes described earlier in “The subpanel DATA” MUST be strictly followed

2) Each term containing a wave specification, e.g. factor 1 or RESW[total] as well as at the end of each criterion line MUST end with a single quotation mark ‘.

3) Each term MUST start with either a space (the case for the first term) or some mathematical symbols, e.g. “(”, “+”, “-”, “^”.

This is essential, as the evaluation of each term has the following architecture:
any symbol’-“term”-“any symbol”-“quotation mark” (termination)

If those rules are not followed, SoFi will fail when treating the individual strings and this causes a bad evaluation of the criterion or unspecific IGOR error messages. Therefore, make sure you know the criteria syntax before changing and defining new manual criteria from the criterion line.

A combined criterion

The user has the possibility to combine two and more criteria into a new one. This is important when for example one criterion is not enough to capture the desired expression. This is typically the case, when, e.g. various time or class settings (subpanel ‘Select based on class / time’) should be combined.

The combined criterion has the following syntax: criterion X’ + criterion Y’

where X and Y stand for the positions of the two criteria. This example shows the sum, but any mathematical combination as well as combinations between normal and GRAND criteria are in theory possible.

Example monitoring C-value runs with the combined criterion

Data from two different instruments, e.g. AMS and PTR-MS are combined and relative rescaling is performed using the C-value approach (see 6.2.4.2). Ideally, the scaled residuals for each instrument should be comparable (same mean value and spread), i.e. their histograms should have maximal overlap. This can be monitored using the following scheme of criteria as example:
criterion 0: \(\text{histogram}(-1,1,10,\text{RESW[total\_matrix]'} \)’ \) (for class 1)
criterion 1: \(\text{histogram}(-1,1,10,\text{RESW[total\_matrix]'} \)’ \) (for class 2)
criterion 2: \(\min(\text{criterion 0'}, \text{criterion 1'}) \)’ \) (using option sum)

For the criterion 0 and 1 the user must select the proper class from the time & class subpanel.

The criteria approach on unconstrained factors

In case many PMF runs with several unconstrained factors are run and the solutions are not inspected manually, the problem arises that the positions of the factors will most likely vary.

Example with 4 factors and 2 PMF runs

PMF run 1

<table>
<thead>
<tr>
<th>position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td>HOA</td>
<td>COA</td>
<td>BBOA</td>
<td>OOA</td>
</tr>
</tbody>
</table>

PMF run 2

<table>
<thead>
<tr>
<th>position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td>HOA</td>
<td>BBOA</td>
<td>COA</td>
<td>OOA</td>
</tr>
</tbody>
</table>

Swapping occurred for position 2 and 3 between the PMF runs 1 and 2. The factors need to be repositioned based on certain conditions / criteria, to guarantee that position 1 carries always HOA, position 2 COA, position 3 BBOA and position 4 OOA.

The problem of repositioning factors on available positions is the classic “sampling without replacement” problem that scales with the factorial of the nb. of positions, e.g., for 6 unconstrained positions there are \(6!\) or \(6*5*4*3*2 = 720\) possibilities. SoFi computes all combinations and evaluates all defined criteria. The scores of each criterion are internally weighed by its mean (to give similar weight to all criteria) and subsequently the sum over all scores is built.

Finally, the combination with the highest sum of scores is returned as chosen combination for ordering the factors. This guarantees that the unconstrained factors are score-sorted in the best possible way.

Active versus passive criteria

A criterion used for repositioning an unconstrained factor is a so-called active criterion. A passive criterion, will simply give the score for the current position, without affecting its sorting. This is typically the case, when factors are constrained and the position is known in advance.

The user governs whether a criterion is active / passive directly from the criteria list by selecting the active ones with the mouse. Active criteria are supposed to be highlighted (see Figure 54). Constraining factors has priority in that active criteria defined for constrained information are subordinate and are simply used for monitoring purposes.
Exploring all combinations works fast and well up to 10 unconstrained factors (3'628'800 combinations per run). If 11 and more unconstrained factors are employed, this strategy becomes cumbersome and extremely inefficient. The user should pay attention not to have more than 10 unconstrained factors when using SoFi Pro with the criteria-based selection.

7.3.3.3. Examples on how to set up the list of criteria
Some examples that illustrate how to set up a criterion follow here.

Correlation between the contribution of factor 1 (hypothetical traffic) and NOx (see Figure 55)
1) Under “type” select “point” and “R-Pearson” or “R-Pearson^2”. Note that the box turns green as this is the color-code for the Pearson option.
2) In the subpanel “DATA” select the option “Time series”, “solution” for “y-type” and then the first entry, i.e. factor_1. In addition, select “external” for “x-type” and NOx data.
3) At this stage, the criterion line shows factor_1;ext_NOx'.
4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.
5) The successfully written criterion is reported under “list of criteria”.

Figure 55  “PMF_criteria_define” subpanel with the example of R_pearson between factor1 and NOx data
Ratio between certain hours of the day for factor 2 (see Figure 56)

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Time series”, “hours” for “y-type” and then select “factor_2[11].”

3) At this stage, the criterion line shows factor_2[12]’.
   The user can directly add the remaining information, i.e. /((factor_2[9]’+factor_2[10]’)/2).
   This expression will evaluate the ratio between the peak at noon compared to the average value of 9 and 10 am.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

![Figure 56](image) "PMF_criteria_define" subpanel with the example of the ratio between the lunch peak to some morning hours for factor 2.
Monitoring the mass fraction of m/z 60 (f60) of the third factor (see Figure 57)

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Time series”, “solution of var.” for “y-type” and then select “factor_3 60:60. In addition, select also “fraction”.

3) At this stage, the criterion line shows factor_3[60].

   This expression will evaluate the mass fraction of m/z 60 of the third factor.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

![Figure 57](image)

“PMF_criteria_define” subpanel with the example of f60 for the third factor.
Monitoring the correlation value R upon a multilinear regression (see Figure 58)
The linear regression is between the time series of factor 1 (hypothetical traffic) and factor 3 (hypothetical biomass burning) to black carbon EC, assuming only traffic and biomass burning to contribute to black carbon.

1) Under “type” select “point” and “Multilinear regression”. Note that the box turns orange as this is the color-code for the Multilinear regression option.

6) In the subpanel “DATA” select the option “Time series”, “solution” for “y-type” and then select “factor_1. In addition, select “external” for “x-type” and BC data.

2) At this stage, the criterion line shows \texttt{factor\_1'; ext\_EC'}

   The user can directly add the remaining information, i.e. \texttt{+factor\_3'} to the first part of the string. The final expression reads: \texttt{factor\_1'+factor\_3'; ext\_EC'}

   This expression will perform a multilinear regression and find the optimal values for the slopes a and b in: \(EC = a \cdot \text{factor}\_1 + b \cdot \text{factor}\_3\). EC is the dependent variable and factor\_1 and factor\_3 are the two independent variables. The current option allows the user to add up to 5 independent variables and always only one dependent variable.

3) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

4) The successfully written criterion is reported under “list of criteria”.

\textbf{Figure 58} “PMF\_criteria\_define” subpanel with the example of the multilinear regression between EC and factor\_1 and factor\_3
Monitoring the ratio between the fraction of m/z 57 and m/z 55 for the second factor (see Figure 59)

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Profiles”, “solution” for “y-type” and then select “factor_2 57:57.

3) At this stage, the criterion line shows factor_2[57]. The user can directly add the remaining information, i.e. /factor_2[55]. This expression will evaluate the ratio between m/z 57 and m/z 55 for the second factor.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

Figure 59  “PMF_criteria_define” subpanel with the example of the ratio of m/z 57 to m/z 55 for the second factor.
Correlation on the diurnal cycle between the contribution of factor 1 (hypothetical traffic) and NOx for the weekdays only (see Figure 60)

1) Under “type” select “point” and “R-Pearson” or “R-Pearson^2”. Note that the box turns green as this is the color-code for the Pearson option.

2) In the subpanel “DATA” select the option “Time series”, “solution” for “y-type” and then the first entry, i.e. factor_1. In addition, select “external” for “x-type” and NOx data.

3) At this stage, the criterion line shows factor_1;ext_NOx.

4) Select diurnal cycle and weekdays from the “Select based on class / time” panel

5) Clicking on “Write” will write this criterion and pass the selected information that is used for the evaluation.

6) The successfully written criterion is reported under “list of criteria”. Note that the criterion is bold, as a temporal sub-selection from the panel “Select based on class / time” is applied.

Figure 60 “PMF_criteria_define” subpanel with the example of R_pearson between factor1 and NOx data for a specific temporal sub-selection.
Monitoring the ratio between the mass fraction of \( m/z \) 44 for the fourth factor to the scaled residual of \( m/z \) 44 (see Figure 61)

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Time series”, “solution of var.” for “y-type” and then select “factor_4 44:44.

3) At this stage, the criterion line shows \text{factor}_4[44]'.
   The user can directly add the remaining information, i.e. \text{RESW}_[44]'.
   This expression will evaluate the ratio between the mass fraction of \( m/z \) 44 of the fourth factor that is supposed to highly contribute to \( m/z \) 44 and the scaled residual of \( m/z \) 44.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

![Figure 61](image)

**Figure 61** “PMF_criteria_define” subpanel with the example of the mass fraction of \( m/z \) 44 of the fourth factor to the scaled residual of \( m/z \) 44.
7.3.3.4. Evaluate the criteria

Pressing on the button “Evaluate list of criteria” allows SoFi to loop over all PMF runs belonging to the current HDF file and to evaluate the user-defined list of criteria (see Figure 62). In addition, the dropdown menu under “criteria file” controls saving/loading options of the list of criteria.

“NEW” creates a new criteria list. If the current list is not empty, it will be deleted.

“LOAD list of criteria” A criteria list stored as an IGOR txt file can be imported
“LOAD list of criteria and evaluation” A criteria list including its scores (makes only sense if the criteria file to be loaded belongs to the same HDF file) stored as an IGOR txt file can be imported

“SAVE list of criteria” A criteria list will be saved in an IGOR txt file
“SAVE list of criteria and evaluation” A criteria list including its scores will be saved as an IGOR txt file

![Image](image_url)

Figure 62 “PMF_criteria” showing the list of possible user-defined criteria.

7.3.3.5. Inspect criteria score

SoFi reports the scores for each criterion and PMF run in the 2nd tab “2) Select PMF run(s)” of the “PMF_criteria” panel (see Figure 64). The default case reports the scores for all PMF runs as is, i.e. the PMF runs are unsorted. Using the adjacent option subpanels allows the user to control various aspects of the scores to better judge which PMF runs are worth being further investigated as a whole.
The button “pop batch table” will open a table which contains all the information on what was selected in the previous tab for the selected criterion (see Figure 63). Note that the first 8 entries in the batch file are reserved and the batch file is not removed, as long as the experiment stays imported into IGOR and the criteria are not removed. Therefore, the user can use the cell of the 8th position to enter user-specific notes, if required.

Figure 63  Batch table for criterion number 3. The first column contains the batch file with information on the criterion. The remaining columns represent the effective time series, its time series index, the effective variables and its index, respectively.

The subpanel “DATA”

“criterion” defines which criterion should be investigated
“sorting” scores of criteria can be “unsorted”, sorted in “ascending order” or “descending order” (see below for more details)
“slope” the slope of the criterion curve can be added or removed from the score graph (see below for more details)
“2nd score” the score of the 2nd classified score after having ordered the unconstrained factors (see below for more details)
“p-values” p-values for the correlation coefficients (Pearson and Kendall tau) can be added or removed from the score graph. This information
The task of the user is to select, using the marquee tool, scores (PMF runs) that should be investigated as full PMF result. To better visualize the range and density of the scores, the score plot can be sorted either in ascending or descending order. When e.g. high scores are expected, the user can more easily select the proper scores in a sorted plot.
Figure 65  Criterion f43 for the profile of factor 5 (gray line) including the slope of the score plot (red curve). Drawing the marquee and select the points around a region of interest highlights them in blue.

The crucial question is where to stop with the selection. When thresholds are already known a priori or derived from other statistical analyses (see stability of criteria 7.3.3.7 or BS 7.3.3.9), they could be directly used within the score plot for the selection. If nothing is known, the user is supposed to investigate the score range by e.g. analyzing packages of PMF runs and to identify the region in the score plot, where the factor solution stops being meaningful/interpretable. As first metric, strong decreases of the score plot suggest a high degree of mixing within the factors. Hence, adding the slope of the score curve (red line) helps in showing the trend of the curve (see Figure 65). Note that the user can select PMF runs to be further investigated using the marquee tool (just presented) or by directly passing the limits on the subpanel (see red rectangle on Figure 65). ‘Left’ and ‘right’ refer to the boundaries on the x-axis. These could be either the index values, e.g. 0 and 100 for left and right, respectively. In case a rolling PMF run is performed, then the user can also pass the time information, e.g. 21/04/2019 and 21/06/2019 for left and right, respectively. In addition, the user can also filter out score points based on the p-values from significance tests as well as the 2nd classified scores values (see 7.3.3.7 for more details).

Once the regions in the score plots for all criteria are chosen the runs fulfilling all selections are identified and highlighted for further analysis “Find PMF runs fulfilling selection for all criteria” and “Apply selection” (see Figure 66). After pressing on “Apply selection”, the user is also asked whether a quick scan over the selected PMF runs should be performed to estimate the Q-statistics and the amount of missing days, a relevant quantity for rolling PMF (see 6.5) runs (see Figure 67).
Figure 66 PMF runs fulfilling the selection for all criteria are visualized and selected for further analysis.

7.3.3.6. Inspect selected PMF runs
The selection (see Figure 66) together with the effectively chosen PMF runs can be visually inspected from the third tab “3) Inspect selection”. The user can compare the score histogram of all PMF runs (gray histogram), with the user-selection (blue histogram) and the effectively chosen PMF runs (green histogram) (see Figure 67). This visualizes which score range is effectively chosen, as the chosen region of PMF runs from the 2nd tab (see Figure 65) sets only the boundaries. However, the effective score limits may be even smaller.

Figure 67 All PMF runs (gray histogram), user-selected PMF runs (blue histogram) and the effective PMF runs (green histogram).

Available options on the adjacent subpanel
“position of criterion” defines which criterion should be investigated
“nb. of bins” defines the number of bins for the two histograms (total PMF runs and effective selection of PMF runs)

“Q-statistics” Q statistics on the effective PMF runs after a quick scan over the PMF runs

“missing days statistics” statistics on the missing time-points (days) after a quick scan over the PMF runs using the daily resolution (automatically defined)

7.3.3.7. Inspect criteria stability

The degree of stability for the score-choice in SoFi can be judged by comparing the highest score of another factor, independent from whether a factor was constrained or not. If the score of another factor is higher than for the current factor (where that score was supposed to be highest), this solution could be regarded as mixed and discarded. In addition, to the score values, its p-values from significance tests can be inspected too (only when checkbox “comparison with all other factors” is checked, see Figure 52). This option will perform a T-test for comparing two regular waves, e.g. the explained variation or a relative lunch peak or a Chi-Test for a correlation coefficient. If the p value is low, the scores tend to be statistically significant (see Figure 68). Note that for factors with higher score values higher than for that under consideration, an arbitrary p-value of one is passed pointing towards mixed/bad PMF solutions.

![Figure 68](image-url)

Criterion score explained variation of m/z 60 for factor 3 (gray line) including the score of highest other factor (black) and the p-values from a T-test (black line). Mainly only the gray points at the beginning of the plot are statistically significant and should be considered. For all the other points, the score of the other factors is mostly higher and these PMF runs should be discarded, as they are mixed.
7.3.3.8. Different number of factors in one HDF file

This option is relevant when inspecting multi-factor solutions with the criteria panel, e.g. PMF runs within one HDF file with e.g. five and six factors. When defining a criterion for the sixth factor, the score plot for this criterion will be empty half of the time, namely for all five-factor runs. This will cause all five-factor PMF runs to be missing in the final solution as they have a non-defined value (NaN) and they would be all discarded when finding the PMF runs that fulfill all criteria selections. Therefore, SoFi passes a zero to these non-defined values allowing the user to select these runs (zero values in Figure 69).

![Score plot for the sixth factor in a rolling PMF with five- and six-factor solutions. The score curve is unsorted (left) and sorted in descending order (right). Note half of the score points have a-value of zero. These points represent PMF runs with five factors and therefore no score was evaluated for the sixth factor, as it is inexistent.](image)

Moreover, inspecting specific factor-runs from a multi-factor solution, e.g. the correlation coefficient between a factor and its tracer for 5-factor solutions only out of all PMF runs, is difficult on the entire score plot. The user can specify which factor-solution should be extracted (red rectangle in Figure 70 a)). The resulting score plot is visualized for these runs allowing the user to make selections based on score values for a specific factor-solution, e.g. score values for 5-factor PMF runs only. Finally, all user-selections are stored and the score plot over all PMF runs reports all selections. In the discussed example the blue points in Figure 70 b) are the selected 5-factor PMF runs. They are alternated by the non-selected gray points that represent the 6-factor PMF runs.

SoFi prevents averaging PMF runs with different number of factors. Whenever the user selects with the marquee or from the button “Apply selection” (Figure 65) some PMF runs, then PMF runs with unequal number of factors within the same PMF window or in case of rolling for the same rolling window (6.5), are automatically disabled. (In case a selection of PMF runs in an HDF
file with multi-factor solutions leads to overlap period, SoFi will only consider the factor solutions with the highest occurrence. When averaging PMF runs from various HDF files, then the factors with the same name specification (see 6.3.1) are grouped and averaged together.

Figure 70  
(a) score plot for 5-factor PMF runs only. The user has selected all PMF-runs within the marquee (blue points).
(b) score plot for all factor-solutions. The previously selected 5-factor PMF runs are the blue points, whereas for the same period marked in (a) the gray points are the non-selected 6-factor PMF runs. All other gray points laying outside of the marquee in (a) are 5- and 6-factor PMF runs, respectively.
7.3.3.9. **Bootstrap runs on the base case**

In case the ratio between a PMF result and its tracer, proxy values or correlation coefficients are known, that is typically the case when a base case already exists or the user has some knowledge over seasonal PMF runs when performing a rolling PMF (see 6.5), then bootstrapping (see 6.4.1 for details on the bootstrap mechanism) this statistic allows to assess its variance. This information is essential, as the resulting variance can be directly used as boundaries in the respective criterion plot and avoiding a tedious data analysis for the assessment of the criteria boundaries by the user.

Such types of bootstrap analyses can be directly performed using the pink panel (see Figure 71)

**Options for the first tab**

- **“criterion line”**
  - the current criterion is visualized. In this example the correlation coefficient between factor 1 and EC_traffic is monitored. Note that factor_1 received the prefix _ext_ too. This means that the wave factor_1 coming from the base case MUST be stored in the same folder where EC_tr resides, i.e. in the external reference folder.

- **“index”**
  - index wave of factor_1, as factor_1 might contain less points compared to the default PMF run (see blacklisting data under 5.4) The user is supposed to manually extract the wave “tseries_index” from the Parameter folder of the relevant result file.

- **“define BS blocks”**
  - crucial data points that drive the statistic must be always present (see blocked bootstrap analysis). The user can create blocks for the bootstrap runs within this section. Points belonging to the same block will be bootstrapped separately and the number of points per block is preserved. The user can enter directly the index of points that should belong to a block or make use of the “select based on class / time” panel for the selection.

- **“Visualize BS blocks”**
  - user-defined blocks are reported in the table below. If no blocks are defined, all data belongs to the block “all” and each point can be resampled individually.
Run the bootstrap analysis
As soon as the BS blocks are defined, the user can define the number of resamples and perform the BS runs. For small datasets, in the order of thousands of points, performing 100 BS runs takes a few seconds.

Inspect the BS results
The third tab allows the user to explore the BS results and to extract the information of the spread of the statistic (interquartile range Q25 to Q75 or the 10th to 90th percentile range of the data).
Note that the number of resamples is a free parameter and should be investigated by the user. The goal is to assess the number of resamples for which the spread of the statistic doesn’t show substantial differences anymore (see Figure 73).

Figure 73  3rd tab of the bootstrap analysis panel for the assessment of the variance of a criterion with the BS results for 100 resamples (left panel) and 1000 resamples (right panel).

In this example the IQR or the 10th and 90th percentile don’t really vary between 100 and 1000 bootstrap resamples, suggesting that already 100 bootstrap runs are sufficient for assessing the spread of the statistic under consideration. Here, the 10th percentile of 0.49 could be directly used as lower boundary in the correlation coefficient criterion. This offers the enormous advantage that the user does not need to explore the criteria ranges anymore, as the boundaries arise from a statistically analysis. In addition, the bootstrap analysis can be performed to any defined criterion without restrictions, as long as the statistic is known. Moreover, bootstrapping is more powerful compared to the p-value analysis for correlation coefficients, as it aims to directly assess the variance of the statistic. These boundaries are more realistic in that they differentiate between the “good” and the “less good” values, whereas the p-values distinguish rather between the “bad” and the “less good” correlation coefficients.
7.3.1. **File settings and logfile**

The user can pop the main settings of the PMF call using the button “Details: file settings & logfile” (Figure 74). The log file shows the history of all PMF runs belonging to the current PMF call (Figure 75).

![Figure 74](image1.png)

**Figure 74** Main file settings of the actual PMF call.

![Figure 75](image2.png)

**Figure 75** Log file of the current PMF call, here “sensitivity_II”.

7.3.2. **Analyze chosen PMF run(s)**

The selected PMF run(s) under 7.3.1 is/are imported into IGOR and the main results are evaluated after pressing on “III.b Analyze chosen PMF run(s)” (Figure 47). The following parts are executed:

- Importing the results
- Calculating for each PMF run the residual matrices, explained variation matrix, relative matrices over profile and time series for the profile (contribution-weighted) and the time series
7.3.2.1. **Averaging**
SoFi Pro allows the user to average over several PMF runs. This is controlled via the option “avg. PMF run(s) (red rectangle on Figure 76).

**Available options:**

- “no average”: only single PMF runs are imported and evaluated
- “compute avg. (save single runs)”: single PMF runs and average over these runs is computed and stored in SoFi. This option is relevant for a small number of PMF runs only
- “compute avg. (only average)”: average of PMF runs is computed and stored in SoFi only. This option is relevant when averaging over many runs that would cause an “out of memory” in IGOR when also stored as single PMF runs in IGOR

7.3.2.2. **Averaging and data options**
The user can control the type of resolution for the result. The default case contains the average for “raw” time resolution. “Hourly”, “daily”, “weekly”, “monthly” and “yearly” resolution can be added and will be performed for each single PMF run and also for the average over the single runs. Moreover, the user can control which data should be evaluated in SoFi (see previous list and Figure 76).

**Note that averaging over lots of PMF runs could cause the averaging process to be very slow. When disabling the checkbox “evaluate quantiles” from the averaging option “data stored of” list, the quantile statistics is not evaluated. The result will contain the average and the standard deviation only and lowers the computational time by a factor 2-3.**

**Note that when averaging over bootstrap runs (see 6.4.1) the repeats per run are first removed.**
Factor specification, already discussed under 6.3.1) can be accessed for the current HDF file also from the “Select_PMFruns” panel using the button “factor specs”. Moreover, factors can also be added, e.g., SV-OOA and LV-OOA can be combined to OOA on a new row in the “factor specs” wave (Figure 77). After averaging the PMF run(s), these combined factors can be used in the red result panel for further analysis.

Note that the combined factors MUST follow the list of real factors, i.e., they MUST be at the end of the list.
factor specs option for the modification of the factor names. Combination of two and more factors are also possible.

7.4. Information on the result folder

The next lines describe the structure of the result folders, the two main parameter files that keep all main settings of the current PMF call as well as the information stored in the result folder of a single PMF run.

7.4.1. Structure of the result folders

root:SoFi:Results

".Variables" All variables pertinent to the main result panel. The user must not change, delete or add anything in here.

".Overview" The waves relevant for the averaged results (see 7.3.2)

".Final" This folder contains the final time series and profile matrices as mean and standard deviation, as well as median and quantiles. This data is useful for post-SoFi analysis, e.g. wind and trajectory analysis using ZEFIR (Petit et al., 2017)

".Plot" plotted waves

".Graph_macro" macros for the recreation of the plots from the result panel tabs
The result folders are listed here, e.g. sensitivity_I, abbreviated as sens_I hereinafter (see Figure 78)

The parameter files (general data and error matrix, time series, variable information, parameter files, convergence criteria for ME-2, name strings, unit string, log file, group information)

All representations relevant to the PMF call (see 7.3.1)

Contains the individual PMF runs of a PMF call, if these runs have been selected for the analysis (see 7.3.1)

---

**Figure 78** Data browser showing the Result subfolders, in particular for the “sensitivity_I” PMF call.

### 7.4.2. Parameter files

There are two main parameter files that declare all relevant information of the PMF call, solution_para and batch_file. Both are stored under root:SoFi:Results:[RUNNAME]:Parameter:.

#### 7.4.2.1. Solution para

Solution_para is a text wave with 17 positions reserved for various general information:

- “solution_para[0]” free position
- “solution_para[1]” normalization during the PMF iteration
- “solution_para[2]” robust mode and its threshold value
- “solution_para[3]” threshold value for the missing data
- “solution_para[4]” C3 value
- “solution_para[5]” weight model type including the parameters
- “solution_para[6]” weight application type
- “solution_para[7]” CO2-related weight
- “solution_para[8]” mass closure
- “solution_para[9]” relative weighing (C-value)
“solution_para[10]”  fpeak value
“solution_para[11]”  a-value type and a-values of constrained information
“solution_para[12]”  reserved for pulling type
“solution_para[13]”  averaged S/N formula
“solution_para[14]”  cellwise S/N formula
“solution_para[15]”  reserved for statistical error propagation
“solution_para[16]”  rolling mechanism

7.4.2.2. Batch file
Batch file is a numerical matrix containing the information for every PMF run.

batch_file[number of PMF runs][varying, depending on the nb. of constraints]

- batch_file[0][0]  nb. of factor
- batch_file[0][1]  iteration index for a fixed nb. of factors
- batch_file[0][2]  rolling index
- batch_file[0][3]  total index, i.e. row dimension of the batch_file
- batch_file[0][>4]  information of constraints (over time and/or profile). Each constraint (time series, profile) receives a separate column.

7.4.3. Main information
The user defines which PMF runs should be considered for the analysis and upon pressing on “III.b Analyze chosen PMF runs(s)” the single PMF runs are imported (see 7.3). The result folder has the following name, e.g. for the PMF call sens_I and the first PMF run:

“/[HDF filename]:Solutions:run_0”

where the zero stands for the run nb., i.e. the first run for this case. A result folder contains typically the following information:

- “curr log”  log file of the current PMF run (log file of ME-2)
- “Input data”  Shows the PMF input for the current PMF run. This is relevant for e.g. the bootstrap strategy (6.4) or the rolling mechanism (6.5).
- “ini_file”  Text wave containing the instruction file passed to ME-2.
- “fpeak_matrix”  Relevant, if fpeak runs have been performed

The next result waves are stored in the corresponding resolution folders, i.e. “raw”, “hourly”, “daily”, “weekly”, “monthly” and/or “yearly”.

- “sol_ts/pr”  are the PMF results
- “pr/ts_constr”  Contains the information of the a-value constraints. It is a 3D matrix with nb. of factors in y and the lower and higher a-value limit in x
for the first two layers. The third layer contains the α-value for every constrained model entry.

“REL_pr/ts” Relative matrix for the time series and the profile matrix. The relative profile matrix is contribution-weighted.

“RES_mx” residual matrix
“RESW_mx” scaled residual matrix
“RESABS_mx” absolute residual matrix
“RESWABS_mx” scaled absolute residual matrix
“Q_mx” Q-matrix
“EXPL_mx” Explained and unexplained variation matrix factor with the x over time, y over the variables and z over the nb. of factors.

“RES_ts/pr” averaged residual matrix over time / profile for all type of RES
“EXPL_tseries” Explained and unexplained variation over time
“EXPL_profile” Explained and unexplained variation over profile
8. SoFi – PMF run(s) analysis

8.1. Select a PMF run

The PMF_analyze panel (Figure 79) displays the results that have been calculated when pressing on “III.b Analyze chosen PMF run(s)” (Figure 47), and hence it reacts very fast upon changes. The panel is divided into three main tabs from which the various graphs can be plotted.

“General information” contains four sections “Select PMF run for exploration”, “Info for all PMF run(s) in adjacent list”, “settings for chosen PMF run...” and “Extract PMF solution”

“Overview plots” Factor time series, profile as well as overview residual plots (absolute residuals, scaled residuals, absolute scaled residuals, Q)

“Detail plots” Fraction plots, factor plots, scatter plots, residual analysis, residual histogram, correlation matrix

“Averaged solution” Plots for evaluating the PMF error, time-dependent factor profiles, time-dependent a-values

![PMF_analyze panel](image)

**Figure 79** “PMF_analyze” panel – tab “General information”.

-106-
8.2. General information

8.2.1. Select PMF run for exploration

The user selects the PMF run (one run at the time) that should be inspected.

When nothing is selected (pressing SHIFT on the selection removes the current selection) and if “compute average” is chosen (see 7.3), the averaged solution is presented as a result.

8.2.2. Info for all PMF run(s) in adjacent list

This part shows some overview statistics of the averaged PMF runs, in case “compute average” is chosen (see 7.3).

8.2.3. Settings for chosen PMF run from adjacent list

“consult log ...” pops the log file written by ME-2 for the current run.
“Swap factor position...” If swapping occurred for unconstrained or loosely constrained factors, the user has the possibility to redefine the order of the factors. Swapping affects only the current/chosen PMF run in IGOR (Figure 80).

![Swapping subpanel showing the old positions (left column) and the new positions (right column).](image)

8.2.4. Extract PMF solution

The set of environmentally reasonable PMF runs, defined as final PMF solution can be extracted and stored under root:SoFi:Results:Final. This option is currently disabled and will be enabled and improved very soon.
8.3. Subpanels graph and data

Once the PMF result panel (red panel) is popped two subpanels “graph” and “data” are popped too.

“Graph panel” contains options related to the graph, e.g. mean or median representation, mode of representation (line, point, cross, etc.), size of representation, etc.

“Data panel” contains options related to the type of data, e.g. time series or diurnal cycle

These subpanels (Figure 81), once popped, remain constantly available, even if the user changes tab. Changing tab will load them automatically with the new options for the current graph. The same subpanels were already presented in 5.3.1 for the data, 5.3.2 for time-dependent options and 5.3.3 for the graph options and the reader is referred to those sections to learn more about these subpanels.

![Figure 81](image_url)

The two subpanels “graph” and “data” loaded with the settings for the PMF factor profiles (left) and time-dependent options for the factor time series (right).

8.4. Overview plots

Factor time series, profiles and the overview residuals can be visualized. Figure 82 shows some possible graphs that are governed from this tab. In addition, temporal information can be represented as cycle plots, i.e. daily, weekly, monthly and yearly cycle.
Figure 82  Factor profile (top left) and residuals over the variables (top right). The residuals over the variables are represented by the mean and interquartile range (IQR). Factor contributions (bottom left) and daily cycle residuals (bottom right) represented with the median, the bars for the IQR and the whiskers for the 10th to 90th %.

For constrained information, e.g. factor profiles Figure 82, the shaded area spans the area between lower and higher constrains with the center of the area representing the anchor (constrained profile). A comparison between this gray area and the resulting profile (colored bar) shows in which direction (up or down) the solution was pulled during the in ME-2 run.

The spread of the average solution, e.g. standard deviation or quantile statistics, as shown for the residuals shown in Figure 82 can be controlled from the subpanel “GRAPH”. Switching between the various resolutions “raw”, “hourly”, “daily”, “weekly”, “monthly” and “yearly” resolution is controlled from the subpanel “DATA”.

-109-
8.5. Detail plots

8.5.1. Fraction

Fractional representation in form of bar and pie charts is available over time and profile for contribution and (un)explained variation data. The user can choose between bar and pie chart representation. Figure 83 shows some possible graphs from the “fraction” tab.

![Figure 83](image)

**Figure 83** bar plot of daily cycle for variable m/z 55 (top left), (un)explained variation as pie chart over a user-defined episode (top right), contribution-weighted bar plot of profile plot (bottom left) and pie chart of mass-weighted variable m/z 60 over the various factors (bottom right).
8.5.2. Standard

The standard tab allows plotting the information over time (time series plot) or over the profile (variable plot) for a multitude of user-selected tracers. Figure 84 shows some possible graphs from the “standard” tab.

The user can add several traces form the current PMF run or from any other PMF run, by simply going back to the selection tab, selecting another PMF run and adding to the existing graph.

Figure 84 daily cycle of factor 1 (hypothetical traffic) and its tracer EC (top left), weekly cycle of factor 1 from several PMF runs (top right), time series of factor 3 (hypothetical BBOA) with its proxy m/z 60 (bottom left) and profile of factor 1 with an adjacent external traffic profile (bottom right).
Moreover, plots generated from the “Overview plots” or the “Fraction” tab can be loaded into the “Standard” tab. This allows the user to add other traces to the existing graphs for a better visualization and comparison. Choose either “overview plot” or “fraction plot” under the option win temp. for adding the plot from the overview or fraction plot. Pressing either on the tab “Standard” or on the “New” button will finally add the preexisting plot (Figure 85). Select the trace to be added from the available options and then press on the “Add” button for adding this trace. IMPORTANT: the option “win temp. “must be set to current, otherwise the new trace cannot be added (Figure 86).

**Figure 85** Plot a preexisting plot from the “Overview plots” or “Fraction” tab.

**Figure 86** Add a new trace from the “Standard” tab to a preexisting plot.
8.5.3. Scatter

The scatter tab allows plotting the information as a scatter plot and color-coding it based on other information over time (time series plot) or over the variables (variable plot) for a multitude of user-selected tracers. Figure 87 shows some possible graphs from the “scatter” tab.

The user can compare traces between the current PMF run or also between various PMF runs, using the checkbox “hold” for the y-axis, x-axis or color-axis, respectively. This keeps the trace constant to the previously selected PMF run.

Figure 87  factor 1 (hypothetical traffic) vs. its tracer EC color-coded by its proxy m/z 55 (top left), factor 4 vs. factor 4 from another PMF (top right), factor 3 (hypothetical BBOA) vs. the explained variation of its proxy m/z 60 color-coded by time (bottom left) and factor 3 (hypothetical BBOA) vs. an external BBOA profile (bottom right).
8.5.3.1. **Option model subtraction**
SoFi Pro offers the possibility to compare a specific modeled outcome to the measured quantity that the model should ideally reproduce. In the case of ACSM data with primary and secondary organic aerosol sources the experimentalist can e.g. remove the primary contribution of specific variables and compare the modeled fractions of the secondary sources (typically the entries of the factor profile) to the secondary measured contribution. Consult Canonaco et al., (2015) for more details on its application.

**Example**
The fraction of m/z 44 and 43, i.e. $f_{44}$ and $f_{43}$ of the modeled secondary factors should be compared to the contribution of measured $f_{44}$ and $f_{43}$ from the secondaries only.

y-type and x-type must read “model subtraction”. y-var and x-var are m/z 44 and m/z 43, respectively and the option “fraction” should be checked. The table on the right on Figure 88 contains the group names for each factor, here POA and SOA and “actual type” is the name of the group of factors that will be used for the comparison. The contribution of all factors that do not belong to the group “actual type” is automatically removed from the total measured contribution on the variables y-type and x-type.

**Figure 88** Model subtraction panel with the first three factors belonging to the group “POA” and the last three to the group “SOA”. The factors being further analyzed are those defined under “actual type”, i.e. SOA.
Given the linear relationship in PMF, the variability of the measured points (red points in Figure 89) is supposed to be captured by the linear combination of the modeled factor profiles (green points in Figure 89). This implies that the measured points that are properly reproduced by the PMF model are those that are on a line if e.g. two factor profiles are modeled. This is the case here, as light and dark green points represent two SOA factors form a rolling PMF run (see 6.5 for details on the rolling strategy).

Figure 89  Outcome of the model subtraction option for f44 vs f43. The POA contribution of f44 and f43 has been removed leading to the red points. The light and dark green points are the fraction of the SOA factor profiles.

Note that when performing “model subtraction” plots on absolute values, the modeled fraction of the factor profiles cannot be compared to the measured contribution. Hence, only the measured variables after having subtracted the contribution from the factors that should not be further inspected are plotted.

Moreover, inspecting the shape of contribution subtracted measured points (red points in Figure 89) might also reveal information on how many factors should be used for the PMF model. If for example these red points lie on a line, only two factors are enough to capture their variability. For the case where the shape of these points is triangular, then three points are required. If the shape of the points is more spherical-like, then only one factor suffices.
8.5.4. Residuals

This tab hosts features relevant to the residual matrix. On one hand, the entire (scaled) residual matrix is accessible. On the other hand, residual histograms can be drawn for the total as well as for single variables or classes, if previously defined.

8.5.4.1. Entire residual matrix

The entire (scaled) residual matrix can be visually inspected. The adjacent slider that controls the color-code of the matrix allows to dynamically inspect various regions of the (scaled) residual matrix and to verify that the entries of the matrix are random, i.e. unstructured over time and over the variables throughout all ranges (see Figure 90).

![Figure 90](image)

Subsection of the scaled residual matrix color-coded between -0.4 and +0.4, based on the user-defined slider “SCAN of residual matrix” value on the subpanel “GRAPH”.

8.5.4.2. Residual histogram

Residual histograms are an easy and fast method for investigating whether the PMF result contains some systematic under- or overestimation (see Figure 91 for some examples). Ideally, the PMF result doesn’t contain any systematic scaled residuals, i.e. the scaled residuals considering all input cells scatter around zero in a noisy way. This leads to a unimodal distribution, ideally Gaussian-shaped (the internally performed Chi squared statistical test tells the user the likelihood to the Gaussian-curve).

Following histograms are accessible
“residual” the residual only is taken. This reveals how good the data is modeled. The shape is supposed to be unimodal. It is not expected to be Gaussian-like, as PMF doesn’t directly fit the residual, but the scaled residual (see below)

“abs. residual” absolute residual histogram. The histogram will be on the positive quadrant only.

“scaled residual” scaled residual histogram. This quantity reveals how well the PMF model fitted the data. Ideally, the probability density function is unimodal and Gaussian-like. In addition, according to Paatero and Hopke (Paatero and Hopke, 2003), the scaled residual histogram is supposed to mainly scatter between +/- 3.

“abs. scaled residual” absolute scaled residual histogram. The histogram will be on the positive quadrant only.

“Q residual” Q histogram. The histogram will be on the positive quadrant only.

Figure 91 Histogram for total scaled residual (left) and residual of variable m/z 44. Note the extreme high bars at +/- 4 due to the robust mode. These lines are not considered for the Gaussian fit and its subsequent statistical significance test.

If the actual residual varies from this shape in that it is bi- or multimodal or showing some high spikes at negative or positive sides, further investigation of the PMF model is required. It might be that the number of factors is too high/low, if constraints were applied that they don’t match the natural variability of the datasets, etc.

Relative weighting of classes (C-value approach)
In case the analyst has performed relative reweighting of the errors of certain classes, their (scaled) residual histograms are plotted separately and proper statistical tests can be performed to judge, whether it is likely that the (scaled) residuals have the same underlying distribution.

-117-
The following statistical analysis is directly accessible from SoFi:

“ANOVA (equal variance)” a one-way ANOVA test with the assumption of equal variance and the underlying normal distribution for all (scaled) residual sampling distribution (classical ANOVA testing with the F distribution or T distribution for two classes). In addition, Bartlett’s test for the equal variance is also computed and the result represented in the adjacent text box.

“ANOVA (unequal variance)” a one-way ANOVA test with the assumption of the underlying normal distribution for all (scaled) residual sampling distribution (Welch test)

“ANOVA (non-parametric)” a non-parametric or rank one-way ANOVA test is performed. (Dunn-Holland-Wolfe test is performed that accounts for various classes with a different number of class members)

Figure 92 dataset split in two classes and errors unweighted (left), the scaled residual of the two classes is optimized using the C-value approach (equilibrate group scaled RES)

Based on the reported p values for each statistical test the analyst can judge how likely the test hypothesis $H_0$ is true, i.e. the mean values of the (scaled) residual distributions for each class arises from the same underlying distribution.

An example is given in Figure 92, where a dataset is artificially split into two classes (class_I in green and class_II in orange) and the errors of class_II were manually increased by 3 orders of magnitude. The left panel shows the PMF run where class_II is underestimated (errors are 1000 times larger than for the variables of class_I), as their scaled residuals are almost all positive. The green and the orange distribution are likely to occur from the same underlying normal distribution with ~24% probability (from a statistical point of view still satisfactory, as typical $\alpha$-values are in the range 5%, 2% or 1%). However, applying the dynamic equilibrate group scaled RES (C-value
approach) made the scaled residuals for both classes even more similar with a p value of ~96%, i.e. both distributions arise from the same underlying normal distribution with 96% probability.

8.5.5. Correlations

Factor time series as well as factor profiles can be correlated to external data and within each other. The following options are available:

- “R_Pearson (squared)” R-Pearson value useful for linear relationships
- “R_Spearman (squared)” also rank correlation, useful for any type of relationship
- “Kendall tau (squared)” rank correlation coefficient that analyzes concordant and discordant pairs
- “R_uncentr (squared)” cosine similarity

For all these options the representation can be either in image or table format (Figure 93). For the Pearson and Spearman correlation coefficients an additional table contains the p-values as well, to better judge on the significance level of the correlation.

![Correlation Image Plot and Correlation Table](image)

**Figure 93** Correlation image plot and correlation table between the factor contribution and external tracers defined under the panel “Data”.

Tables on the subpanel “Data” allow the user to select the proper factors and externals. Default value is that all entries are used for the correlation (Figure 94).

The subpanel “Graph” also contains the critical value for a significance test based on the threshold. The user can vary the threshold on the same panel and can check immediately for acceptable minimum R values. This test is only available within the option R-Pearson and Kendall tau.
Figure 94  Graph (left) and data (right) subpanels for the section “correlation”.

8.5.6.  HR plot

For HR-AMS data analysis the family-stacked version of the factor profile plot is supported in SoFi. The tab “HR-family” is accessible, in case the text vector containing the family information for the variables with the name “speciesfamilytext” is stored in addition to the PMF input data (import data, see 4.1). Moreover, all factors are equipped with the OM:OC, O:C and H:C calculation (Figure 95). The O:C and H:C ratios are given as mole fractions and the user can choose between one of the two options “Aiken ambient” (Aiken et al., 2008) and “Canagaratna ambient” (Canagaratna et al., 2015) method for visualizing these ratios.

Figure 95  HR family stacked plot for HR-AMS analysis.
8.6. **Averaged solution**

This tab is useful for the analysis of the averaged solution in particular for rolling PMF runs. It contains statistical information on the degree of coverage of the PMF input data, a method for quantifying the PMF error, statistics on the time-dependent a-values and time-dependent factor profiles. However, the tab “error estimation” will also serve well when quantifying the PMF error for regular (non-rolling) PMF runs.

8.6.1. **Repeats**

When bootstrapping or performing a rolling PMF run some modelled data points might be absent, e.g. for only a few number of replicates when bootstrapping or very tight thresholds are applied for the criteria. The tab “repeats” reveals the coverage and how good the averaged PMF result represents the PMF input data. In addition, the amount of repeats over time is also plotted (see Figure 96). Ideally, the amount of repeats is uniformly distributed over the entire PMF input and strong differences suggest a still insufficient amount of selected PMF runs, systematic over-/underrepresentation of the data due to a non-ideal number of factors, or constraints, etc.

![Figure 96](image)

“Repeats” tab with some statistics on the PMF run and on the amount of repeated input points. This is particularly useful for bootstrapped or rolling PMF runs.
8.6.1. Mass error estimation

Applying resampling strategies such as bootstrap (see 6.4.1) and techniques aiming to systematically explore the rotational ambiguity such as a-value/ pulling (see 2.4) generate repeats for a timepoint \( i \) in the PMF solution. The variability within these points can be used to infer the rotational and statistical uncertainty. These two types of uncertainties are discussed in the following and are referred to as PMF error. It should be noticed that more types of errors might be added to the final PMF error, such as for example the model error (error of using a different number of factors, a different type of anchor for the a-value constraints) or the error related to the criteria such as the type of criterion and its related thresholds. The proposed PMF error in this study is given by the following formula:

\[
PMF_{error,i} = 100 \cdot \left( \frac{\text{spread}}{\text{mean/median}} \right)_i
\]

where the spread could be the differences between the third and first quartile or the standard deviation for a time-point \( i \). The slope of a linear fit when plotting the spread against the mean or median value represents the relative error. Figure 97 represents both, the scatter plot and the probability density function (pdf) of the top formula for each timepoint \( i \).

Figure 97  
Spread in terms of IQR is plotted against the median value for factor 2 (left) and its corresponding density function plot (right). The mass error for this factor is half the slope in percentage, i.e. +/- 17 \% on the scatter plot. On the frequency plot, the error is more represented by the overall median value of 42.9 \%. Hence, the error is +/- 21 \%, given possible small mean values for some points \( i \) causing extremely large values affecting the overall mean of 57.5 \%.
8.6.2. Employed $\alpha$-values

When performing random $\alpha$-values in e.g. a rolling PMF run this tab visualizes the statistics of the employed $\alpha$-values over time and as histogram (see Figure 98 and Figure 99). This graph reveals information on the degree to which the constrained information required to adjust to the PMF input data in time. High $\alpha$-values with a small spread suggest that for these episodes the constrained information required more adjustment/freedom.

![Figure 98](image1.png)  
**Figure 98** (left) effective $\alpha$-values for the constrained factor profiles over time and (right) histogram representation of these $\alpha$-values.

In addition to the mean, standard deviation, median and quantile statistics, the probability density function (pdf) can be visualized too.

![Figure 99](image2.png)  
**Figure 99** Probability density function of the $\alpha$-values over time.
8.6.1. **t-profiles**

Rolling PMF aims to assess time-dependent factor profiles. This information is shown within this tab by plotting the factor profile over time as an image plot. The profiles can be normalized over the profile (sum of the variables equals one) or over time (each variable is divided by its mean value). Moreover, a 2-way ANOVA test is performed that tests all profiles for independence. The Null hypothesis ($H_0$) describes the assumption that all profiles present in the $t$-profile matrix are drawn from the same distribution. The results are visualized on the subpanel “GRAPH”. (see Figure 100)

![Image](image.png)

**Figure 100**  factor profile in the y-axis plotted over time. Variations in intensity of the variables reveal changes in the profile shape.

The user can also inspect the profiles by plotting averaged profiles over time. Drawing a marquee on the interested region (note that the x-axis only matters, i.e. the entire profile is drawn, even when not all variables (y-axis) are within the marquee, loads the marquee option after pressing on the right mouse button (see Figure 101). This process can be repeated until the desired averaged profiles can be directly compared together (see Figure 102).
Figure 101  factor profile in the y-axis plotted over time including the option for plotting averaged profiles.

Figure 102  two averaged factor profiles including their statistic (top graph for the winter and bottom graph for the summer period).
## 9. List of definitions and abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME-2</td>
<td>Multilinear Engine second version</td>
</tr>
<tr>
<td>OS</td>
<td>Operating System</td>
</tr>
<tr>
<td>PMF</td>
<td>Positive Matrix Factorization</td>
</tr>
<tr>
<td>PMF2</td>
<td>Solver for the bilinear positive matrix factorization algorithm</td>
</tr>
<tr>
<td>PMF3</td>
<td>Solver for the trilinear positive matrix factorization algorithm</td>
</tr>
<tr>
<td>PMF call</td>
<td>Calling ME-2 to solve the PMF algorithm for certain settings, e.g. varying the number of factors, exploring the a-value, etc.</td>
</tr>
<tr>
<td>PMF run</td>
<td>Refers to a single PMF run out of a PMF call</td>
</tr>
<tr>
<td>PMF result / solution</td>
<td>Represents all PMF runs from one or more PMF calls that are environmentally reasonable and must be considered after a thorough rotational ambiguity exploration and the propagation of the statistical uncertainty.</td>
</tr>
<tr>
<td>POA</td>
<td>Primary organic aerosol</td>
</tr>
<tr>
<td>SA</td>
<td>Source apportionment</td>
</tr>
<tr>
<td>SOA</td>
<td>Secondary organic aerosol</td>
</tr>
<tr>
<td>SoFi</td>
<td>Source Finder software</td>
</tr>
<tr>
<td>SoFi Pro</td>
<td>Key that enables the advanced utilities in SoFi</td>
</tr>
<tr>
<td>RT SoFi Pro</td>
<td>Key that enables the automated SoFi software</td>
</tr>
</tbody>
</table>
10. Acknowledgment

Special thanks are due to Dr. Jay Slowik, Dr. Yulyia Sosedova, Dr. Carlo Bozzetti, Dr. Kaspar Dällenbach for a critical reading of this manuscript during the past few years.

A special thank goes to Prof. Dr. André Prévôt, Prof. Dr. Urs Baltensperger and Prof. Dr. Alexander Wokaun for having made all this possible.

The authors would like to thank Dr. Pentti Paatero who was always available for technical and scientific questions during the learning process of ME-2 solver and its script code.

Thanks also go to all PSI and non-PSI students and PhDs, who tested and reported inconsistencies, suggestions, etc. to the author during these years.
11. Literature


