Spring 1-1-2011

Characterization of Positive Matrix Factorization Methods and Their Application to Ambient Aerosol Mass Spectra

Ingrid Marie Ulbrich
University of Colorado at Boulder, ulbrich@colorado.edu

Follow this and additional works at: https://scholar.colorado.edu/chem_gradetds
Part of the Applied Mathematics Commons, Atmospheric Sciences Commons, and the Environmental Sciences Commons

Recommended Citation
https://scholar.colorado.edu/chem_gradetds/35

This Dissertation is brought to you for free and open access by Chemistry & Biochemistry at CU Scholar. It has been accepted for inclusion in Chemistry & Biochemistry Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Characterization of Positive Matrix Factorization Methods and Their Application to Ambient Aerosol Mass Spectra

By

Ingrid Marie Ulbrich

B.S., Massachusetts Institute of Technology, 2000

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
Department of Chemistry and Biochemistry
2011
This thesis entitled:

Characterization of Positive Matrix Factorization Methods and Their Application to Ambient Aerosol Mass Spectra

written by Ingrid Marie Ulbrich

has been approved for the Department of Chemistry and Biochemistry by

____________________________________
Jose-Luis Jimenez

____________________________________
Michael P. Hannigan

Date: ______________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Atmospheric aerosol has impacts on health, visibility, ecosystems, and climate. The organic component of submicron aerosol is a complex mixture of tens of thousands of compounds, and it is still challenging to quantify the direct sources of organic aerosol. Organic aerosol can also form from a variety of secondary reactions in the atmosphere, which are poorly understood. Real-time instrumental techniques, including the Aerosol Mass Spectrometer (AMS), which can quantitatively measure aerosol composition with high time and size resolution, and some chemical resolution, produce large volumes of data that contain rich information about aerosol sources and processes. This thesis work seeks to extract the underlying information that describes organic aerosol sources and processes by applying factor analytical techniques to organic aerosol datasets from the AMS. We have developed a custom, open-source software tool to compare factorization solutions, their residuals, and tracer-factor correlations. The application of existing mathematical techniques to these new datasets requires careful characterization of the precision in the data and the factorization models’ behavior with these specialized datasets. We explore this behavior with synthetic datasets modeled on AMS data. The synthetic data factorization has predictable behaviors when solved with “too many” factors. These behaviors then guide the choice of solution for real aerosol datasets. The factor analyses of real aerosol datasets are useful for identifying aerosol types related to sources (e.g., urban combustion and
biomass burning) and secondary atmospheric processes (e.g., semivolatile and low-volatility oxidized organic aerosol). We have also factored three-dimensional datasets of size-resolved aerosol composition data to explore the variability of aerosol size distributions as the aerosol undergoes processing in an urban atmosphere. This study provides evidence that primary particles are coated with condensed secondary aerosol during photochemical processing, shifting the size distribution of the primary particles to larger sizes. Application of these three-dimensional factorization techniques to other complex aerosol composition datasets (e.g., that use thermal desorption or chromatography for further chemical separation) has the potential to yield additional insights about aerosol sources and processes.
for my parents

who always tell me that
I can accomplish anything I set my mind to,
but it might take longer than one day

and for my grandparents

who always tell me
I shouldn’t work too hard
and I should go home before it gets dark out
Acknowledgements

Back when I had a “real job” (with a real salary!) I worked with a lot of people who had Ph.D.’s. I saw that they knew how to approach and solve big problems in a way that I didn’t, and I wanted to learn how to do that. While going back to school was not my immediate first choice, I knew that it would be necessary to gain this experience. Furthermore, I realized that getting a Master’s degree would not challenge me with a big enough project to get everything I wanted. And so I embarked on the journey to get a Ph.D. Thank you to Praveen Amar, Rawlings Miller, Gary Kleiman, and John Graham for your excellent advice about choosing a program and an advisor.

My seven years of graduate study at CU have been a fantastic opportunity to work with a lot of really good scientists who are also really good people. This work would not have been possible without the AMS and the many people behind its development and the early success at analyzing its complex datasets. I acknowledge those who began this work and developed it to the point that my analyses were possible: John Jayne, Doug Worsnop, James Allan, Ravi Alfarra, and Qi Zhang. Likewise, I am thankful that I did not have to write the algorithms that solve the factorization models used in this work! Thank you to Pentti Paatero for developing these tools and making them available. Early discussions with Pentti Paatero and Phil Hopke about the application of PMF to our datasets were very helpful.

I must also acknowledge the past members of the Jimenez group for breaking in a new professor and beginning his early training, for their assistance in preparing me for my oral exam by telling me to stop talking, and for incorporating my analyses into their own work. Donna Sueper has been a wonderful resource for programming the PET, including giving me coding
advice, always knowing a little Igor trick, and taking over the code for a year so, which allowed me to focus more on research. Of course, this work would not have been possible without the assistance of my advisor, Prof. Jose-Luis Jimenez. I have always appreciated that we have a common connection to Cambridge, we usually speak the same language about work and science, and that I became the “anonymous student” who got you in trouble with The Boss for staying at work too late.

The first four years of this work culminated with the text that forms Chapter 2 of this thesis. Many thanks to Manjula Canagarata for her extensive revision of an early draft of that work. It not only improved the paper significantly, but taught me a great deal about manuscript organization and structure. I have tried to pass those lessons on when reading drafts of others’ work. Much of the analysis for that manuscript was done and most of the draft was written at Folsom St. Coffee, where they know how I like my chai and my tea.

Many thanks to family, not just during my graduate career. My parents have always encouraged and supported my endeavors, acknowledged my frustrations when I did not achieve my goals immediately at the outset of a project, and waited patiently for me to realize when I had achieved success. I am so glad that Jose was at CU so that I could come home after ten years in Boston. Thank you to my grandparents, who think I work too hard, and listen patiently every time I tell them that’s the way it needs to be sometimes. Thank you to Michelle, Eric, and Rachel, who were ready teach my niece and nephew to call me “Dr. Auntie Ingrid” seven years ago. And finally, thank you to my fiancé Brian, who not only loves and puts up with me, but in especially in the last several months has also been my chauffeur, personal chef, thesaurus, manual of style, and empathetic therapist. Don’t worry, I’ll cook dinner tomorrow.
# Table of Contents

## Chapter 1
### Introduction
- 1.1 Aerosol Overview ......................................................... 1
- 1.2 Identification of Targets for Particle Controls ......................... 2
- 1.3 Chemometrics Overview .................................................. 3
  - 1.3.1 Chemometric applications to ambient aerosol ................. 3
- 1.4 Framework for This Thesis ............................................. 5

## Chapter 2
### Interpretation of Organic Components from Positive Matrix Factorization of Aerosol Mass Spectrometric Data
- 2.1 Chapter Introduction ..................................................... 7
- 2.2 Methods ................................................................. 13
  - 2.2.1 Aerosol Mass Spectrometer (AMS) ......................... 13
  - 2.2.2 Factorization Methods ........................................... 15
    - 2.2.2.1 Positive Matrix Factorization (PMF) ............. 15
    - 2.2.2.2 Singular Value Decomposition (SVD) .......... 22
  - 2.2.3 Data Sets ......................................................... 22
    - 2.2.3.1 Real Pittsburgh Dataset ............................ 22
    - 2.2.3.2 Synthetic Datasets .................................... 25
  - 2.2.4 Statistical Comparisons of Mass Spectra .................... 27
    - 2.2.4.1 Reference Spectra ....................................... 27
    - 2.2.4.2 Statistics of Correlation .............................. 28
- 2.3 Results ................................................................. 29
  - 2.3.1 Real Pittsburgh Data ............................................ 29
    - 2.3.1.1 Solutions as a Function of Number of Factors .... 30
    - 2.3.1.2 Rotations .................................................. 41
  - 2.3.2 Synthetic AMS Data ............................................ 45
    - 2.3.2.1 Solutions of Synthetic Data Base Cases ........ 45
    - 2.3.2.2 Separation of Correlated Factors ................. 50
- 2.4 Discussion ............................................................. 53
- 2.5 Chapter Conclusions .................................................. 59

## Chapter 3
### Three-dimensional factorization of size-resolved organic aerosol mass spectra from Mexico City
- 3.1 Chapter Introduction ..................................................... 62
- 3.2 3-Dimensional Matrix Factorization and its Application in the Literature ......................... 66
  - 3.2.1 Mathematical Techniques for 3D Matrix Factorization ...... 66
  - 3.2.2 Research Reporting 3-Dimensional Factorizations Using Particle Size Information .... 71
- 3.3 Methods ................................................................. 73
  - 3.3.1 Mexico City Measurements During the MILAGRO Field Campaign .... 73
  - 3.3.2 AMS Sampling of Particle Time-of-Flight Data ............ 74
Tables

Chapter 3
Table 3.1 Summary of research that applied 3D factorization techniques to datasets of size-resolved aerosol chemical composition............................69
Table 3.2 Types of factors identified in each solution family from 50 “seed” trials of the 3-vector and vector-matrix models solved by ME-2……………...95

Appendix A
Table A.1 Parameters of the four sources for the synthetic dataset………………..170

Appendix D
Table D.1 Correlations between PMF factor and tracer time series..................243
Table D.2 Correlations between PMF factor and reference mass spectra............244
Table D.3 Correlations between input factors in the 3-factor synthetic datasets…..245

Appendix H
Table H.1 Downweighting multipliers for m/z’s with interference from late-arriving air molecules at m/z 29 and m/z 44.........................................................299
Table H.2 Summary of research that applied 2D factorization techniques to aerosol size distributions or size-resolved aerosol composition datasets………300
Table H.3 Details of research that reports application of 2D factorization techniques to aerosol size distributions......................................................301
Table H.4 Details of research that reports application of 2D factorization techniques to aerosol size distributions and particle-phase aerosol composition.....303
Table H.5 Details of research that reports application of 2D factorization techniques to aerosol size distributions and gas-phase composition............304
Table H.6 Details of research that reports application of 2D factorization techniques to aerosol size distributions and gas-, and particle-phase composition...305
Table H.7 Details of research that reports application of 2D factorization techniques to size-resolved aerosol composition datasets..............................306
Table H.8 Details of research that reports application of 2D factorization techniques to size-resolved aerosol composition datasets in which the sizes are combined into one matrix.........................................................308
Table H.9 Configuration of the computer and ME-2 algorithm used to solve the factorizations..........................................................309
Figures

Chapter 2

Figure 2.1  Schematic of PMF factorization of an AMS dataset .......................... 11
Figure 2.2  Screenshot of the PMF Evaluation Tool ........................................... 17
Figure 2.3  Values of Q/Qexp and the maximum of RotMat ................................ 30
Figure 2.4  Mass fraction of each factor .............................................................. 31
Figure 2.5  Factors from the three-component PMF solution of the real Pittsburgh dataset ................................................................. 33
Figure 2.6  Uncentered correlation between spectra from the AMS Spectral Database and solution spectra ................................................... 34
Figure 2.7  Time series of factors from the real Pittsburgh dataset and Q/Qexp ...... 39
Figure 2.8  Distributions of scaled residuals for each m/z for the three-component solution of the real Pittsburgh dataset ................................. 40
Figure 2.9  Q/Qexp vs. FPEAK for the real Pittsburgh and synthetic datasets ......... 41
Figure 2.10 Three-factor solutions of the real Pittsburgh dataset for selected FPEAK values ......................................................................................... 42
Figure 2.11 Correlations of PMF factors to each other as they change with FPEAK .. 44
Figure 2.12 Correlations between input factors for synthetic datasets ................. 51
Figure 2.13 Correlation versus FPEAK between PMF factors and reference mass spectra and time series ............................................................. 57

Chapter 3

Figure 3.1  Schematic representation of factorization methods of 3D matrices ...... 68
Figure 3.2  Schematic representation of data collection for chemically-resolved particle size-distribution data ......................................................... 76
Figure 3.3  Values of Q/Qexp for the 3-vector and vector-matrix factorization models with 1 to 8 factors ................................................................. 93
Figure 3.4  The best 3-vector solution, which has four factors, solved with ME-2 ... 99
Figure 3.5  Values of Q/Qexp for four-factor solutions of the vector-matrix model in which the factor mass spectra have been constrained to a priori spectra vs. the constraint relaxation parameter, β ................................................. 101
Figure 3.6  Residuals and Q/Qexp values for solutions of the constrained vector-matrix model with four factors ......................................................... 105
Figure 3.7  The best constrained vector-matrix solution, which has four factors .... 107
Figure 3.8  Diurnal average size distributions of OOA, HOA, BBOA, and LOA from the best solution of the constrained vector-matrix model .......... 109
Figure 3.9  Comparison of the factors from the best solutions of the 3-vector and vector-matrix models and the HR-MS solution .......................... 112
Figure 3.10 Comparison of average size distributions as calculated from the 3-vector model, vector-matrix model, and tracer method for OOA, HOA, BBOA, and LOA ............................................................... 117

Appendix A

Figure A.1  Fit of QUAIL ROOST II data by several normalization schemes....... 160
Figure A.2  Selected spectra from the synthetic dataset ....................................... 171
Figure A.3  Non-rotated results of the positive matrix factorization of the synthetic dataset…………………………………………………………………..172
Figure A.4  Rotated results of the positive matrix factorization of the synthetic dataset…………………………………………………………………..173

Appendix C
Figure C.1  Schematic diagram for calculation of error matrices for real and synthetic AMS data……………………………………………..220
Figure C.2  Comparison of errors calculated for the real and synthetic datasets……..221

Appendix D
Figure D.1  Comparisons of the uncentered correlation with Pearson R…………..225
Figure D.2  Solutions of the real Pittsburgh dataset in robust and nonrobust modes.226
Figure D.3  PMF solutions of the real Pittsburgh dataset with 2, 4, and 5 factors…..227
Figure D.4  Pearson R between spectra from the AMS Spectral Database and solution mass spectra…………………………………………….228
Figure D.5  Diurnal profiles of OOA-2, tracers, and residuals………………………229
Figure D.6  Solutions of the real Pittsburgh dataset from 64 random starts……….230
Figure D.7  Results from bootstrapping analysis of the real Pittsburgh dataset……231
Figure D.8  Time series of factors and residuals of the real Pittsburgh dataset……232
Figure D.9  Histograms of scaled residuals of selected m/z;s………………………233
Figure D.10 Change in Q/Qexp in the real Pittsburgh dataset vs. factor time series…236
Figure D.11 Pearson R of PMF factors to each other as they change with FPEAK…237
Figure D.12 PMF solutions of the 2-factor synthetic base case with 2–5 factors…..238
Figure D.13 PMF solutions of the 3-factor synthetic base case with 2–5 factors…….239
Figure D.14 Three-factor solutions of the 3-factor synthetic dataset for selected FPEAK values………………………………………………………240
Figure D.15 Pearson R between input factors for synthetic datasets…………………241
Figure D.16 Pearson R versus FPEAK between PMF factors and reference mass spectra and time series………………………………………..242

Appendix E
Figure E.1  Correlation of factor mass spectra and time series for two-factor synthetic input datasets and the rotation of their solutions with FPEAK………..247
Figure E.2  Retrieval of the input factors in the two-factor synthetic datasets……..249

Appendix F
Figure F.1  Q/Qexp vs. number of factors in solution and Q/Qexp vs. FPEAK for m/z 43………………………………………………………………………..250
Figure F.2  Fraction of variance and mass concentration for solutions with 1 to 6 factors for m/z 43………………………………………………………250
Figure F.3  Solution with 1 factor for m/z 43………………………………………..251
Figure F.4  Solution with 2 factors for m/z 43………………………………………..251
Figure F.5  Solution with 3 factors for m/z 43………………………………………..251
Figure F.6  Solution with 4 factors for m/z 43………………………………………..251
Figure F.7  Solution with 5 factors for m/z 43………………………………………..252
Figure F.8  Solution with 6 factors for m/z 43……………………………………………………….252
Figure F.9  Q/Qexp vs. number of factors in solution for m/z 46………………………………….253
Figure F.10 Fraction of variance and mass concentration for solutions with 1 to 6 factors for m/z 46……………………………………………………………………………253
Figure F.11 Solution with 1 factor for m/z 46……………………………………………………….254
Figure F.12 Solution with 2 factors for m/z 46……………………………………………………….254
Figure F.13 Solution with 3 factors for m/z 46……………………………………………………….254
Figure F.14 Solution with 4 factors for m/z 46……………………………………………………….254
Figure F.15 Solution with 5 factors for m/z 46……………………………………………………….255
Figure F.16 Solution with 6 factors for m/z 46……………………………………………………….255
Figure F.17 Q/Qexp vs. number of factors in solution for m/z 64………………………………….256
Figure F.18 Fraction of variance and mass concentration for solutions with 1 to 6 factors for m/z 64…………………………………………………………………………….256
Figure F.19 Solution with 1 factor for m/z 64……………………………………………………….257
Figure F.20 Solution with 2 factors for m/z 64……………………………………………………….257
Figure F.21 Solution with 3 factors for m/z 64……………………………………………………….257
Figure F.22 Solution with 4 factors for m/z 64……………………………………………………….257
Figure F.23 Solution with 5 factors for m/z 64……………………………………………………….258
Figure F.24 Solution with 6 factors for m/z 64……………………………………………………….258

Appendix H
Figure H.1  Particle time-of-flight (PToF) data at m/z 43 and m/z 16……………………………..287
Figure H.2  Fraction of points in the 3D matrix whose error was increased to a minimum value of 1 ion…………………………………………………………………………….288
Figure H.3  Average signal-to-noise ratio (SNR) of each m/z at each particle size………………289
Figure H.4  Mass size distribution for the best solution of the constrained vector-matrix model………………………………………………………………………………………..290
Figure H.5  Case study event on 24 March 2006………………………………………………………291
Figure H.6  Q/Qexp values for 50 seed solutions of 4-6 factors grouped into families with similar factors……………………………………………………………………………292
Figure H.7  Solution of the 3-vector model with 4 factors and the lowest Q/Qexp values…………………………………………………………………………………………293
Figure H.8  Solution of the 3-vector model with 4 factors that include two with similar mass spectra……………………………………………………………………………294
Figure H.9  Five-factor solution of the constrained vector-matrix model…………………………295
Figure H.10 Schematic representation of 2D factorizations of datasets that include size distribution information or size-resolved aerosol composition………………296
Figure H.11 Time to calculate each solution of the 3-vector and vector-matrix models with 1–8 factors using the PMF3 or ME-2 algorithm……………………………………297
Figure H.12 Four-factor solution of the unconstrained vector-matrix model, including a factor with a mass spectrum dominated by m/z’s 67, 81, and 95………298
Chapter 1

Introduction

1.1 Aerosol Overview

Fine particles have important effects on health, visibility, ecosystems, and climate. Most of these impacts are considered to be harmful, and many societies have tried to control particulate emissions in order to reduce their impacts. However, controls can only be effective when the sources of particles and the processes that transform them in the atmosphere are understood. Controls can be applied efficiently if they are targeted toward the particles that have the greatest concentrations or impact. Thus there is a need to understand the relationship between aerosol sources and the processes that transport and transform them in the atmosphere.

Particles can be categorized in relation to their origin in several different ways. Particles can be characterized by how they enter the atmosphere: as primary particles, directly emitted to the atmosphere; or as secondary particles, formed by chemical reactions in the atmosphere. Particles can also be characterized by their sources. Typical sources of primary aerosol include combustion processes (industrial, vehicular, food cooking, and biomass burning, Engel-Cox and Weber, 2007), mechanical processes (brake wear, erosion, mineral dust, sea salt, Engel-Cox and Weber, 2007; Amato et al., 2009), and biogenic emissions (pollen and plant debris, MatthiasMaser and Jaenicke, 1995). These particles can have complex chemical compositions that include metals and metal oxides, organic and elemental carbon compounds, and ionic salts (e.g., Amato et al., 2009). In contrast, secondary particles can form from a host of gas-phase reactions that produce condensable products. These reactions include gas-phase inorganic
reactions that produce ammonium nitrate and ammonium sulfate, gas-phase organic reactions that result in low-volatility products that condense onto the surfaces partition into the volume of existing particles (Pankow, 1994; Robinson et al., 2007), and aqueous-phase reactions that produce species that stay in the particle phase (secondary organic aerosol, or SOA, Altieri et al., 2006). Furthermore, the sources of particles can be categorized as natural or anthropogenic. While each primary source can be neatly assigned as natural or anthropogenic, the categorization of secondary species is more complex because the reactants may both be natural, both anthropogenic, or a combination of natural and anthropogenic (Weber et al., 2007). Certainly, humans will have the most success at controlling anthropogenic particles and precursors. Finally, particles can be classified by their size, with the major classes being PM$_{10}$ (particles with diameters less than 10 $\mu$m), PM$_{2.5}$, and PM$_{1}$ (also known as submicron aerosol). These size classes have been identified by both their natural modes in the atmosphere and their health impacts.

1.2 Identification of targets for particle controls

Using particle source-receptor relationships as the basis for controls on aerosol has been a regulatory practice for more than 30 years (Cass, 1978). Aerosol controls have focused on primary anthropogenic particles. In the United States and Europe, controls have mainly been required for large stationary sources and motor vehicles because these sources have large contributions to PM$_{10}$ and PM$_{2.5}$ and control technology has been relatively simple and available for these sources (DMU, 2004). Indirect controls on secondary inorganic aerosol precursors (mainly NO$_x$ and SO$_2$) have also been implemented, but US restrictions on emissions of these molecules have been only moderately successful at reducing the formation of these aerosol and decreasing their impacts (Tsimpidi et al., 2008). Thus, many areas within these regions still
suffer from unsafe levels of particulate air pollution. In addition, we do not understand all of the climate impacts of aerosol. Submicron aerosol have disproportionate impacts on health (Yeh et al., 1996) and climate (Rasch et al., 2008); in addition, submicron aerosol mass is dominated by secondary inorganic and organic particles (Zhang et al., 2007; Jimenez et al., 2009). Therefore there is a need to better understand the sources and processes of secondary, submicron aerosol. A good understanding of secondary aerosols requires the identification of precursor compounds, their reaction schemes, and kinetics, and all of these should be known before effective controls can be proposed. This is especially true of SOA, for which the number of possible precursor and product compounds is vast. This understanding could be aided by quantitative tools that relate measurements of aerosol chemical composition to sources. Ideally, these tools should explain the contribution from and temporal variation in particles with important contributions to submicron OA.

1.3 Chemometrics overview

Chemometrics is a subfield of analytical chemistry that applies mathematical tools and models to chemical data to identify patterns that can be used to elucidate chemical information (Massart et al., 1988). In general, multidimensional techniques are applied to chemical datasets to identify underlying properties of datasets and reduce large datasets to a smaller number of descriptors that are a meaningful characterization of the original data. In fact, a recent issue of Science had a special section about the challenges of “too much data” (2011), and an article on climate data states that “more resources need to be dedicated to the development of sophisticated software tools for sifting through, accessing, and visualizing… observed data…” (Overpeck et al., 2011). Chemometrics has several subfields, including pattern recognition, optimization and control, and multilinear analyses such as multiple linear regression and factor analysis.
1.3.1 Chemometric applications to ambient aerosol

Several techniques from the subfield of pattern recognition have been applied to atmospheric particle data since the 1960’s (Blifford and Meeker, 1967). The techniques used to identify sources of ambient aerosol from two-dimensional datasets all solve the bilinear unmixing model in which each factor is composed of two vectors. One vector represents the factor’s chemical composition, and the other represents that factor’s concentration in each sample. These models assume that aerosol measured at a receptor comes from multiple sources, and the chemical composition of the bulk aerosol is the linear sum of the particles from all the sources.

Four main algorithms/approaches have been used to solve this model: principle component analysis (PCA, Pearson, 1901), chemical mass balance (CMB, Miller et al., 1972), positive matrix factorization (PMF, Paatero and Tapper, 1994), and UNMIX (Henry, 2003). These techniques differ in the requirement for a priori knowledge and assumptions about the orthogonality of the chemical composition of the sources, and these differences have implications for choosing and interpreting the solutions of the models. PMF requires no a priori information about the chemical composition of the aerosol and the factor chemical composition and mass concentrations must be positive. For these reasons, many modelers choose PMF for a more open-ended analysis that produces physically-realistic factors.

These mathematical techniques have traditionally been applied to chemically speciated datasets in which the chemical composition includes concentrations of a combination of elements, ions, and EC, OC (Engel-Cox and Weber, 2007; Reff et al., 2007; Viana et al., 2008). CMB techniques have also been applied to datasets of organic markers for known sources (Schauer et al., 1996). The chemical data for these studies often represents trace portions of the
aerosol, not the total aerosol mass. Most samples have low time resolution (24 hours, and often not sampled every day).

The sources identified from these studies include many types of combustion aerosol (from industrial stationary sources, vehicle emissions, biomass burning/woodsmoke, and meat cooking) and mechanically generated aerosol (from mineral dust, road dust, sea salt, others) (Engel-Cox and Weber, 2007; Reff et al., 2007; Viana et al., 2008). The sources mainly represent primary aerosol, and sometimes secondary inorganic aerosol. The factorization results can be used by air pollution agencies to justify controls on sources.

In contrast, secondary organic aerosol is not usually apportioned in these studies. The resolution of the organic chemical information is either very fine but very selective (markers); CMB analysis of organic markers usually quantifies only sources based on small number of tracers. In contrast, a second approach to including organic aerosol composition in source apportionment analyses, the inclusion of OC bins from the IMPROVE protocol, is coarse and not very selective. The OC bins are operationally defined based on temperature and carrier gas composition when combusting carbon, not chemically defined to identify organic compounds (Chow et al., 2001).

1.4 Framework for this thesis

At the outset of this thesis work, there was a deficiency in understanding the complex composition of bulk organic aerosol. We needed (and still need) to understand in greater detail the contribution of primary vs. secondary organic aerosol, and how the organic aerosol is chemically transformed in the atmosphere. A new instrument had recently become available that could be used to sample bulk, submicron aerosol quickly and quantitatively, but with low
speciation for organic molecules (Jayne et al., 2000; Jimenez et al., 2003). AMS datasets have inherent internal correlations because of molecular fragmentation patterns in electron impact mass spectrometry.

Preliminary studies for understanding the main components of the aerosol were being developed based on tracers from the mass spectra. These studies began to give insight into the split of primary vs. secondary OA. But we preferred a technique that did not require a priori knowledge of tracers for identifying factors, and began working with PMF. However, the AMS data is very different from the data traditionally used with PMF. All of our data comes from the same instrument, and it has strong internal correlations. Thus we wanted to characterize the behavior of this technique with this new type of data.

The thesis work has three main parts. First, we developed software to run the algorithm, display the results, and show useful statistics and comparisons to additional data. Second, we characterize PMF results for 2-dimensional AMS datasets using real and synthetic AMS organic dataset. Third, we extend the factorization to 3-dimensional models applied to size-resolved chemical data from the Time-of-Flight-AMS.
Chapter 2

Interpretation of Organic Components from Positive Matrix Factorization of Aerosol Mass Spectrometric Data

2.1 Chapter Introduction

Fine particles have important effects on human health, visibility, climate forcing, and deposition of acids and nutrients to ecosystems and crops. Organic species represent an important fraction of the submicron aerosol at most locations (Kanakidou et al., 2005; Zhang et al., 2007a). Typically referred to as organic aerosols (OA), they are the sum of multiple primary and secondary sources that can evolve due to aging processes. Apportioning organic aerosols into their sources and components correctly is a critical step towards enabling efficient control strategies and model representations.

The organic source apportionment problem has been approached by several techniques. Turpin and Huntzicker (1991) utilized the ratio between elemental carbon and organic carbon (EC/OC) from filter samples to estimate primary and secondary OA. Schauer et al. (1996) used molecular markers with a chemical mass balance (CMB) approach to apportion OA extracted from filters and analyzed by GC-MS. Several sources with unique markers can be identified, but source profiles must be known a priori, sources without unique markers are not easily separated, and only primary OA sources are identified. Szidat et al. (2006) have separated anthropogenic and biogenic OA based on water solubility and $^{14}$C/$^{12}$C ratios and found a major biogenic influence in Zurich, Switzerland. The technique has very low time resolution (many hours to

---

1 Published as Ulbrich et al., 2009.
several days) and can identify only a few categories of sources. Traditional OA filter measurements suffer from low time resolution (several hrs. to days) and positive and negative artifacts (Turpin et al., 2000).

The last 15 years have seen the development of a new generation of real-time aerosol chemical instrumentation, most commonly based on mass spectrometry or ion chromatography (Sullivan et al., 2004; DeCarlo et al., 2006; Williams et al., 2006; Canagaratna et al., 2007; Murphy, 2007). Current real-time instruments can produce data over timescales of seconds to minutes and have reduced sampling artifacts compared to filters. Single-particle mass spectrometers (e.g., PALMS, ATOFMS, SPLAT) have used particle classification systems to group particles based on composition or other characteristics (Murphy et al., 2003). A fast GC-MS system (TAG) has been developed that may allow the application of the molecular marker technique with much faster time resolution than previously possible (Williams et al., 2006). However GC-MS as typically applied discriminates against oxygenated organic aerosols (OOA) (Huffman et al., 2009), which is the dominant ambient OA component (Zhang et al., 2007a), and thus may limit the applicability of this technique by itself. It is highly desirable to perform source apportionment based on the composition of the whole OA. This information cannot be obtained at the molecular level with current techniques, however several techniques are starting to characterize the types/groups of species in bulk OA (Fuzzi et al., 2001; Russell, 2003; Zhang et al., 2005a; c).

The Aerodyne Aerosol Mass Spectrometer (AMS) belongs to the category of instruments that seeks to measure and characterize the whole OA. It has been designed to quantitatively measure the non-refractory components of submicron aerosol with high time resolution (Jayne et al., 2000; Jimenez et al., 2003) and produces ensemble average spectra for organic species every
few seconds to minutes (Allan et al., 2004). Several groups have attempted different methods to deconvolve the OA spectral matrix measured by a Q-AMS (Lanz et al., 2007; Marcolli et al., 2006; Zhang et al., 2005a; c; Zhang et al., 2007a). Zhang et al. (2005a) first showed that information on OA sources could be extracted from linear decomposition of AMS spectra by using a custom principal component analysis (CPCA) method applied to OA data from the Pittsburgh Supersite from 2002. The resulting factors were identified as hydrocarbon-like organic aerosol (HOA, a reduced OA) and oxygenated organic aerosol (OOA) and were strongly linked to primary and secondary organic aerosol (POA and SOA), respectively, based on comparison of their spectra to known sources and their time series to other tracers. OOA was found to dominate OA (~2/3 of the OA mass was OOA), in contrast to previous results at this location (Cabada et al., 2004). Zhang et al. (2007a) used the Multiple Component Analysis technique (MCA, an expanded version of the CPCA) for separating more than two factors in datasets from 37 field campaigns in the Northern Hemisphere and found that the sum of several OOAs comprises more of the organic aerosol mass than HOA at most locations and times, and that in rural areas the fraction of HOA is usually very small. Marcolli et al. (2006) applied a hierarchical cluster analysis to Q-AMS data from the New England Air Quality Study (NEAQS) from 2002. Clusters in this data represented biogenic VOC oxidation products, highly oxidized OA, and other small categories. Lanz et al. (2007) applied Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994; Paatero, 1997) to the organic fraction of a Q-AMS dataset from Zurich in the summer of 2005. The six factors identified in this study were HOA, two types of OOA (a highly-oxidized, thermodynamically stable type called OOA-1 that correlates well with aerosol sulfate; and a less-oxidized, semi-volatile type called OOA-2 that correlates well with aerosol nitrate), charbroiling, wood burning, and a minor source that may be influenced by food
cooking. Lanz et al. (2008) applied a hybrid receptor model (combining CMB-style a priori information of factor profiles with the bilinear PMF model) specified by the Mulilinear Engine (ME-2, Paatero, 1999) to apportion the organic fraction of a Q-AMS dataset from Zurich during wintertime inversions, when no physically-meaningful components could be identified by the bilinear model alone. Three factors, representing HOA, OOA, and wood burning aerosol, were identified, with OOA and wood-burning aerosol accounting for 55% and 38% of the mass, respectively. More advanced source apportionment methods based on Bayesian statistics, which output a probability distribution instead of scalars for each element of the source profiles and time series and thus contain information necessary for a statistical evaluation of the uncertainty of the output, are under development (Christensen et al., 2007; Lingwall et al., 2008). Bayesian models can also incorporate prior information in a natural and probabilistically rigorous way by specification of the “prior distribution” for each variable. Bayesian methods are expensive computationally, and the more complex output requires greater review by the researcher. Bayesian methods have not been applied to aerosol MS data to our knowledge.

Of the analysis techniques mentioned above, PMF is the most widely used in the atmospheric research community (Lee et al., 1999; Ramadan et al., 2000; Kim et al., 2003; Xie et al., 1999; Larsen and Baker, 2003; Maykut et al., 2003) and its application to PM has been recently summarized in two separate reviews (Reff et al., 2007; Engel-Cox and Weber, 2007). PMF is a receptor-only, factorization model based on mass conservation which requires no a priori information about factor profiles or time trends. PMF has generally been applied to long-term, low-time-resolution datasets, though there has been a call for greater application of source apportionment techniques to air pollution events to facilitate understanding of specific sources for regulatory purposes (Engel-Cox and Weber, 2007). As shown schematically in Fig. 2.1,
PMF is a bilinear unmixing model in which a dataset matrix is assumed to be comprised of the linear combination of factors with constant profiles that have varying contributions across the dataset. All of the values in the profiles and contributions are constrained to be positive. The model can have an arbitrary number of factors; the user must select the solution that “best” explains the data. This is often the most subjective and least quantitative step of PMF analysis and relies greatly on the judgment and skill of the modeler (Engel-Cox and Weber, 2007; Reff et al., 2007). In addition, mathematical deconvolution of a dataset often yields non-unique solutions, in which linear transformations (colloquially referred to as “rotations”) of the factors are possible while the positivity constraint is maintained. The necessity of choosing a number of factors and a particular rotation often complicates the interpretation of the solutions. As clearly articulated by P. Paatero (manuscript in preparation, Pers. Comm. 2007):

“It is unfortunate that introducing a priori information also introduces some subjectivity in the analysis […] However, the tradeoff is often between a successful albeit subjectively aided analysis and an unsuccessful analysis. […]
subjective decisions must be fully and openly reported in publications. […] Hiding the details of subjective decisions or even worse, pretending that no subjectivity is included in the analysis, should not be tolerated in scientific publications.”

Although the application of PMF analysis to data from the AMS and other aerosol mass spectrometers is relatively new, it is quickly becoming widespread. Thus, a detailed characterization of the capabilities and pitfalls of this type of analysis when applied to aerosol MS data is important. UMR AMS datasets are very large with typically several million datapoints (~300 m/z’s per sample, with ~8000 samples for a month-long campaign with 5 min. averaging) and fragmentation of molecules during ionization gives each mass spectrum strongly interrelated data. AMS datasets differ in two fundamental ways from most atmospheric datasets to which PMF has been applied. The structure, internal correlation between some m/z’s created by significant fragmentation of molecules in the vaporization and ionization processes in the AMS, and precision of AMS data are significantly different from datasets of multiple aerosol components (metals, organic and elemental carbon, ions, etc.) measured by several instruments typically used with PMF in previous studies. The error structure is also more coherent and self-consistent due to the use of data from a single instrument, rather than mixing data from different instruments for which the relative errors may be more difficult to quantify precisely, or that may drift differently, etc.

In this work, we apply PMF to data obtained with the quadrupole Aerosol Mass Spectrometer (Q-AMS) during the Pittsburgh Air Quality Study. Three factors, interpreted as HOA, aged regional OOA, and fresh, semivolatile OOA are reported for the Pittsburgh ambient dataset. The ambiguities associated with choosing the number of factors and their best rotations are reported. In addition, sensitivity analyses are performed with synthetic datasets constructed to retain the inherent structure of AMS data and errors. We explore methods that can inform the
choice of the appropriate number of factors and rotation for AMS OA datasets, as well as investigate the retrievability of small factors.

2.2 Methods

2.2.1 Aerosol Mass Spectrometer (AMS)

The Q-AMS has been described in detail elsewhere (Jayne et al., 2000; Jimenez et al., 2003) and only a brief summary is given here. The AMS consists of three main parts: an inlet system which generates a particle beam, a particle sizing section, and a particle composition analysis section. The inlet focuses submicron particles into a narrow beam. A mechanical chopper allows all particles (“beam open”), no particles (“beam closed”), or a packet of particles (“beam chopped”) to pass to the particle sizing region. Particles impact on a flash vaporizer (600°C) at the rear of the sizing region under high vacuum (~10^{-7} Torr). The vapors from non-refractory components are ionized by electron impact (70 eV) and the resultant positively charged ions are analyzed by the quadrupole mass spectrometer.

The Q-AMS can be operated in any of three modes: mass spectrum (MS), particle time-of-flight (PToF), or jump mass spectrum (JMS). In MS mode, the chopper alternates between the beam open and closed positions while the mass spectrometer scans across m/z 1 to 300. Each e.g. five-minute average is the difference between the total open and closed signals and is the ensemble average mass spectrum of thousands of particles. In PToF mode, the beam is chopped and packets of one or a few particles enter the particle sizing region. Particles achieve size-dependent velocities at the exit of the lens which allows measurement of particle size distributions, but only at ~10-15 selected m/z’s. JMS mode is identical to MS mode except in that only ~10 m/z’s are monitored to maximize signal-to-noise ratio (SNR) (Crosier et al., 2007; Nemitz et al., 2008). We use the MS mode data for this study because it has high signal-to-noise
and contains the full structure of the mass spectra and thus the most chemical information. Each sample is the linear combination of the spectra from all particles and species vaporized during the sample period. If JMS data is available, it may be used to replace the MS m/z’s as the JMS data has much better SNR (Crosier et al., 2007). Preliminary analyses show that PToF data contains significant information that can be exploited by PMF-like methods (Nemitz et al., 2008), however this also introduces additional complexities and it is outside of the scope of this paper.

Newer versions of the AMS include the compact time-of-flight mass spectrometer (C-ToF-AMS, (Drewnick et al., 2005)) and the high-resolution ToF-AMS (HR-ToF-AMS, (DeCarlo et al., 2006)). These instruments operate in MS and PToF modes. Conceptually the MS mode from the C-ToF-AMS produces identical data to those from the Q-AMS, except with higher SNR, and thus the results from this paper should be applicable to PMF analyses of such data. The MS mode from the HR-ToF-AMS contains much additional chemical information such as time series of high resolution ions (e.g., both C₃H₇⁺ and C₂H₃O⁺ instead of total m/z 43) that should facilitate the extraction of PMF components. The first application of PMF to HR-ToF-AMS MS-mode data will be presented in a separate publication (DeCarlo et al., 2009).

The datasets used in this study are comprised of only the organic portion of the AMS mass spectrum measured by the Q-AMS, which is determined from the total mass spectrum by application of a “fragmentation table” (Allan et al., 2004) for removing ions from air and inorganic species. The atomic oxygen to carbon ratio (O/C) for UMR MS can be estimated from the percent of OA signal at m/z 44 (predominately CO₂⁺) in the OA MS (Aiken et al., 2008). Percent m/z 44 is reported here as an indication of the degree of oxygenation of representative spectra.
The time series of inorganic species (non-refractory ammonium, nitrate, sulfate, and chloride) are not included in the PMF analysis and are instead retained for a posteriori comparison with the time series of the factors and for use in their interpretation. It is also of interest to perform the PMF analysis on the total spectrum without removing inorganic species (but still removing the large air signals), however this is outside the scope of this paper.

### 2.2.2 Factorization Methods

#### 2.2.2.1 Positive Matrix Factorization (PMF)

Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994; Paatero, 1997) is a model for solving a receptor-only, bilinear unmixing model which assumes that a measured dataset conforms to a mass-balance of a number of constant source profiles (mass spectra for AMS data) contributing varying concentrations over the time of the dataset (time series), such that

\[ x_{ij} = \sum_{p} g_{ip} f_{pj} + e_{ij} \]  

(2.1)

where \( i \) and \( j \) refer to row and column indices in the matrix, respectively, \( p \) is the number of factors in the solution, and \( x_{ij} \) is an element of the \( m \times n \) matrix \( X \) of measured data elements to be fit. In AMS data, the \( m \) rows of \( X \) are ensemble average mass spectra (MS) of typically tens of thousands of particles measured over each averaging period (typically 5 min.) and the \( n \) columns of \( X \) are the time series (TS) of each \( m/z \) sampled. \( g_{ij} \) is an element of the \( m \times p \) matrix \( G \) whose columns are the factor TS, \( f_{ij} \) is an element of the \( p \times n \) matrix \( F \) whose rows are the factor profiles (MS), and \( e_{ij} \) is an element of the \( m \times n \) matrix \( E \) of residuals not fit by the model for each experimental data point \((E=X-GF)\). A schematic representation of the factorization is shown in Fig. 2.1. The model requires no a priori information about the values of \( G \) and \( F \).
normalize the rows in $\mathbf{F}$ (MS) to sum to 1, giving units of mass concentration ($\mu g/m^3$) to the columns of $\mathbf{G}$ (TS). The values of $\mathbf{G}$ and $\mathbf{F}$ are iteratively fit to the data using a least-squares algorithm, minimizing a quality of fit parameter $Q$, defined as

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

(2.2)

where $\sigma_{ij}$ is an element of the $m \times n$ matrix of estimated errors (standard deviations) of the points in the data matrix, $\mathbf{X}$. In the “robust mode” of the algorithm, outliers ($|e_{ij}/\sigma_{ij}| > 4$) are dynamically reweighted throughout the fitting process so that they cannot pull the fit with weight $> 4$. The $Q$-value reported by PMF is calculated using the reduced weights for the outliers. This scaling makes optimal use of the information content of the data by weighing variables by their degree of measurement certainty (Paatero and Tapper, 1994). Additionally, the values in $\mathbf{G}$ and $\mathbf{F}$ are constrained to be positive, reflecting positive contributions of each factor to the measured mass and positive signal in each $m/z$, respectively. A more detailed explanation of the PMF algorithm is provided in Appendix A.

The bilinear model can be solved by the PMF2 (Paatero, 2007) and multilinear engine (ME) (Paatero, 1999) algorithms developed by P. Paatero, or by custom algorithms developed by others (Lu and Wu, 2004; Lee and Seung, 1999; Hoyer, 2004). Here we use PMF2 because of its robustness and wide use in the research community. Future work will explore more complex models using the ME program. All analyses in this study were done with PMF2 version 4.2 in the robust mode, unless otherwise noted. The default convergence criteria were not modified. Since the output of PMF is very large and evaluating it is very complex, we developed a custom software tool (PMF Evaluation Tool, PET, Fig. 2.2) in Igor Pro (WaveMetrics, Inc., Portland, Oregon). More information on the PET is available in Appendix B.
Figure 2.2. Screenshot of PMF Evaluation Panel developed for general examination of PMF solutions.
It is possible that there may be multiple local minima of the Q function (Paatero, 1997); additional solutions can be explored by starting the PMF2 algorithm from different pseudorandom values known as “seeds” within the PMF algorithm. The PET calls the PMF2 algorithm to solve a given problem for a list of values of $p$ and FPEAK or SEED, stores the results for all of these combinations, and allows the user to rapidly display and compare many aspects of the solution matrix and residuals and to systematically evaluate the similarities and differences of the output spectra and time series with known source/component spectra and tracer time series.

**Choosing the Number of Factors**

The number of factors, $p$, in the real dataset is generally unknown. Choosing the best modeled number of factors for a dataset is the most critical decision to the interpretation of the PMF results. Several mathematical metrics have been used to aid determination of this value. A first criterion is the Q-value, the total sum of the squares of the scaled residuals. If all points in the matrix are fit to within their expected error, then $\text{abs}(e_{ij})/\sigma_{ij}$ is $\sim 1$ and the expected Q ($Q_{\text{exp}}$) equals the degrees of freedom of the fitted data $= mn - p(m+n)$ (Paatero et al., 2002). For AMS datasets, $mn >> p(m+n)$, so $Q_{\text{exp}} \approx mn$, the number of points in the data matrix. If the assumptions of the bilinear model are appropriate for the problem (data is the sum of variable amounts of components with constant mass spectra) and the estimation of the errors in the input data is accurate, solutions with numbers of factors that give $Q/Q_{\text{exp}}$ near 1 should be obtained. Values of $Q/Q_{\text{exp}} >> 1$ indicate underestimation of the errors or variability in the factor profiles that cannot be simply modeled as the sum of the given number of components. If $Q/Q_{\text{exp}} << 1$, the errors of the input data have been overestimated. As additional factors are considered, Q is
expected to decrease, as each additional factor introduces more degrees of freedom that should allow more of the data to be fit. A large decrease in $Q$ with the addition of another factor implies that the additional factor has explained significantly more of the variation in the data and has also been used as a metric for choosing a solution (Paatero and Tapper, 1993). A second metric for choosing a best solution is based on the values of the rotational matrix (RotMat, output by PMF and explained below). Some have used the criterion of a solution with the least rotation (lowest maximum value of Rotmat) as one of several qualitative metrics for making the determination of the number of factors (Lee et al., 1999; Lanz et al., 2007). Many studies have concluded that source apportionment models must be combined with supplementary evidence to choose and identify factors (Engel-Cox and Weber, 2007).

Choosing the best number of factors requires the modeler to determine when additional factors fail to explain more of the variability in the dataset. Note that it is possible for one true factor to be mathematically represented by multiple factors which, in total, represent the true factor (Paatero, 2008a). Consider a case in which two true factors make up the data with no error, such that

$$X = GF$$

where $G = [a \ b]$, the matrix of the time series of the two factors, and $F^T = [s \ t]^T$, the matrix of the profiles of the two factors, and $a$, $b$, $s$, and $t$ are column vectors. If the same dataset $X$ is solved with three factors, an exact solution could be obtained as

$$X = [e \ f \ b] \times [s \ s \ t]^T$$

if $e+f = a$. In fact, a case could be constructed in which two factors reconstruct $b$ instead of $a$, generating a second type of 3-factor solution. More combinations are possible when the same $X$ created with 2 factors is solved with 4 factors, e.g.,
\[
\mathbf{X} = [e f b b] \times [s s u v]^T
\]

(2.5)

where \(e+f = a\) and \(u + v = t\), or

\[
\mathbf{X} = [d e f b] \times [s s s t]^T
\]

(2.6)

where \(d+e+f = a\). We refer to this type of behavior as “splitting” of the real factors, where either the MS or TS from a real factor are split into two new factors. Linear transformations (“rotations”, discussed further in the next section) of these solutions are also possible. A rotation of the three-factor solution shown in Eqn. (2.4) could be represented by

\[
\mathbf{X} = [e f b] \mathbf{T} \times \mathbf{T}^{-1} [s s t]^T
\]

(2.7)

where \(\mathbf{T}\) is a 3x3 non-singular transformation matrix and \(\mathbf{T}^{-1}\) is its inverse and is a valid solution to the PMF model as long as the rotated factor matrices \([e f b] \mathbf{T}\) and \(\mathbf{T}^{-1} [s s t]^T\) all have positive values. Thus the rotated solutions need not necessarily contain repetitions of the factors from the original solution. We refer to this later behavior as “mixing” of the real factors.

**Rotational Ambiguity of Solutions**

Despite the constraint of non-negativity, PMF solutions may not be unique, i.e., there may be linear transformations (“rotations”) of the factor time series and mass spectra that result in an identical fit to the data, such that:

\[
\mathbf{G} = \mathbf{G} \mathbf{T} \mathbf{T}^{-1} \mathbf{F}
\]

(2.8)

where \(\mathbf{T}\) is a transformation matrix and \(\mathbf{T}^{-1}\) is its inverse. A given \(t_{ij} > 0\) would create a rotation by adding the mass spectra and subtracting the time series of factors \(i\) and \(j\), while \(t_{ij} < 0\) would create a rotation by subtracting the mass spectra and adding the time series of factors \(i\) and \(j\). An infinite number of “rotations” may exist and still meet the non-negativity constraint. Note that
orthogonal or “solid body” geometric rotations of the factors are only a subset of the possible linear transformations.

PMF2 does not report the possible values of $T$, but does report the standard deviation of possible values of $T$ as the “RotMat” matrix. Larger values in $T$ imply greater rotational freedom of a solution. Specifically, a larger value of $t_{ij}$ suggests that the $i$ and $j$ factors can be mixed while still satisfying the non-negativity constraint. Diagonal elements of $T$ are always 1, and their standard deviations are therefore 0. RotMat for a one-factor solution is always 0. The value of RotMat as a diagnostic has been debated in the literature (Lanz et al., 2007; Lee et al., 1999; Paatero, 2007), and we explore its use as a qualitative indicator of rotational freedom of a given solution (Sects. 2.3.1.2, 2.3.2.2).

With PMF2, once the approximate best number of factors has been determined, a subset of the rotational freedom of the solution may be explored through use of the FPEAK parameter. FPEAK allows for examining approximate or “distorted” rotations that do not strictly follow Eq. (2.7) and thus produce a higher value of $Q$. Of greatest interest are FPEAK values for which $Q$ does not increase significantly over $Q_{\text{FPEAK}=0}$, since the PMF model (Eq. 2.2) is still satisfied with little additional error. Some researchers recommend exploring a range of FPEAKs such that $Q/Q_{\text{exp}}$ increases from its minimum by e.g. 10% (P.K. Hopke, Pers. Comm. 2007). Solutions reported in the literature generally have an FPEAK value between -1 and +1 (Reff et al., 2007). Not all possible rotations can be explored by varying FPEAK, and the characterization of rotations outside of that realm is a topic of current research (Paatero and Hopke, in preparation 2008 2; Paatero et al., 2002). As stated above, none of these metrics or criteria can unequivocally pinpoint the “best” or “true” solution, but a chosen PMF solution can be justified through comparison of factor profiles with known profiles and comparison of factor time series.
with the time series of tracer species. If several plausible solutions are found, e.g. with different FPEAKs, the differences in the solutions are a good indication of the uncertainty of the PMF solution (Nemitz et al., 2008; Paatero, 2007).

**Uncertainty of the Solutions**

The difficult issue of the uncertainty of the solutions is rarely addressed in PMF studies in the literature (Reff et al., 2007). We address this point in this work in two ways: in a qualitative way by running the PMF algorithm from many different random starting points (SEEDs; Paatero, 2007), and quantitatively with bootstrapping with replacement of MS (Norris et al., 2008; Press et al., 2007a).

### 2.2.2.2 Singular Value Decomposition

In contrast to PMF, the singular value decomposition (SVD) of a matrix produces only one factorization (as in Eq. (2.1)) with orthogonal factors. Starting with the factor that explains the most variance of the original matrix, factors are retained in order of decreasing variance of the matrix to explain enough (usually 99%) of the variance of the original matrix. These orthogonal factors usually contain negative values. SVD is applied to selected data matrices and residual matrices to determine the number of factors needed to explain 99% of the variance of the matrix. The relationship between SVD and PMF is described by Paatero and Tapper (1993).

### 2.3 Data Sets

#### 2.3.1 Real Pittsburgh Dataset

The real Pittsburgh dataset investigated here is the same as that analyzed by Zhang et al. (2005a) with the CPCA method. Versions without pretreatments and with pretreatments applied
(filtering for high-noise spikes, 3-point smoothing of m/z time series, and use of cluster analysis (Murphy et al., 2003) to remove unusual spectra as described in Zhang et al. (2005a)) were analyzed with PMF2. Additional information on the Pittsburgh study can be found in previous publications (Zhang et al., 2005a; b; c; Zhang et al., 2007b; Zhang et al., 2004). The study took place September 7-22, 2002 in Pittsburgh, Pennsylvania as part of the Pittsburgh Air Quality Study (PAQS) at the EPA Supersite. 3199 time-averaged mass spectra (5-10 minute averaging) were collected for m/z 1 to 300. Fragments with plausible organic fragments were retained, leaving 270 m/z fragments. Thirty fragments were removed because they couldn’t have plausible organic fragments, have overwhelming contribution from inorganic or gaseous species, or high instrument background (Zhang et al., 2005a). In addition, organic fragments at m/z 19 and 20 are omitted as the signals at these m/z’s are directly proportional to m/z 44 and have negligible contribution (< 0.05% of the total signal) and therefore do not add new information to the factorization analysis. The remaining matrix had 268 columns (m/z) and 3199 rows (time-averaged mass spectra) with 857,332 data points.

The error values for use with PMF were calculated in five steps. First, the initial error values were calculated by the method of Allan et al. (2003) by the standard Q-AMS data analysis software (v1.41). We recommend that for Q-AMS data, version 1.41 or later of the standard data analysis software is used for estimation the errors for use with PMF, as corrections to the error calculation algorithms have been made from previous versions and error matrices calculated from earlier versions may give different factors because of different weighting. Nonsensical behavior of the factors (MS with one dominant fragment or TS that oscillate between zero and several μg/m³ over 5-minute periods) were observed with this dataset when the error estimates from older versions of the Q-AMS data analysis software, but not when v1.41
was used. Second, a minimum error estimate of one measured ion during the sampling time (equivalent to 11 Hz or 0.12 ng/m$^3$, which reflects the duty cycle used during this campaign) or the average of the adjacent error values is applied to any elements of the error matrix ($\sigma_{ij}$) with values below this threshold by

$$
\sigma_{ij} = \max(\sigma_{ij}, \max(1/t_s, (\sigma_{i-1,j} + \sigma_{i+1,j})/2))
$$

(2.9)

where $t_s$ is the time, in seconds, spent sampling each $m/z$. Third, the 3-point box smoothing applied to the dataset was propagated in the error estimates by summing the error or the 3 smoothed points in quadrature. This has the effect of decreasing the noise estimate by a factor of $\sqrt[3]{3}$. Fourth, we follow the recommendation of Paatero and Hopke (2003) to remove variables (TS of $m/z$ in our case) with signal-to-noise ratio (SNR) less than 0.2 (“bad” variables) and downweight variables with SNR between 0.2 and 2 (“weak” variables) by increasing their estimated error values. For this dataset, no columns are “bad” by this definition and 76 of the higher mass fragments ($m/z$ 167-168, 207, 210, 212, 214, 220-223, 230-238, 240-249, 254-300) are “weak” and their error estimates are increased by a factor of 2. Finally, in order to appropriately weight $m/z$’s 44, 18, 17, and 16 (since the latter 3 peaks are related proportionally only to $m/z$ 44 in the organic “fragmentation table” (Allan et al., 2004) whose inclusion therefore gives additional weight to the strong signal at $m/z$ 44), the error values for each of these $m/z$’s are all multiplied by $\sqrt{4}$ (Nga L. Ng, Pers. Comm. 2008; see Appendix D). There are two sets of $m/z$’s that are directly proportional to only one other $m/z$ ($m/z$ 48 and $m/z$ 62; $m/z$ 80 and $m/z$ 94), but these signals are much smaller than those of the $m/z$ 44-group and the effect of this adjustment is negligible. Note that the order of steps four and five are arbitrary; even after changing the error estimates for $m/z$’s 44, 18, 17, and 16 their SNR’s are approximately 25 and they are not “weak” variables. The downweighting of so many $m/z$’s in the datasets lowers the
calculated Q-values; \( Q/Q_{\text{exp}} \)-values reported in this work have therefore been recalculated by undoing this scaling (but still applying the robust criterion) so that \( Q/Q_{\text{exp}} \)-values are related to the error estimated for each point from the data prior to downweighting.

No adjustments are made to the errors in this analysis to reflect “model error” that may occur because the true factors are not constant as assumed by the PMF model. Increasing the error values to reflect this error may downweight real phenomena that are part of the true data. Note that small negative values that are the result of differences (caused by noise) between the beam open and beam closed measurements in the instrument data are not changed, nor are their corresponding error values altered.

2.2.3.2 Synthetic Datasets

Each synthetic dataset was created by combining selected MS and TS into \( F_{\text{input}} \) and \( G_{\text{input}} \) matrices, respectively, which were then multiplied to form an \( X_{\text{input}} \) matrix (\( X_{\text{input}} = G_{\text{input}} \times F_{\text{input}} \)), the forward calculation of Eq. (2.1) with \( e_{ij} = 0 \).

Synthetic noise was added to the difference spectrum synthetic data, such that the noisy synthetic data, \( x' \), were calculated by

\[
x_{ij}' = (\text{Poisson}(\text{open}_{ij}) - \text{Poisson}(\text{closed}_{ij})) \times \text{CF} + \text{Gaussian}(0, 0.0002)
\]

where random noise is generated from a Poisson random number generator (Igor Pro v6.03) with a mean and variance of the number of ions observed in the open and closed MS of that point (\( \text{open}_{ij} \) and \( \text{closed}_{ij} \), respectively), \( \text{CF} \) is the conversion factor from ions per \( m/z \) per averaging period to \( \mu g/m^3 \), and electronic noise is estimated from a normal Gaussian distribution with a standard deviation of 0.0002 \( \mu g/m^3 \). The amount of 0.0002 \( \mu g/m^3 \) is an estimate of the electronic noise present during periods of low signal in several \( m/z \)’s > 239. Poisson noise is used for ion counting noise instead of Gaussian noise because many of the small signals do not
have sufficient counts to reach a Gaussian distribution to a good approximation. The sum of ion counting and electronic noises represents most of the noise in a Q-AMS dataset, but does not reflect “particle-counting statistics noise” from events when a large particle is vaporized and “extra” (much greater than average) signal is detected at only one m/z during the scanning of the quadrupole across the m/z range (Zhang et al., 2005a).

The synthetic Poisson-distributed error values for these datasets were approximated by a method parallel to the estimation of errors for real data (see description in Appendix C). The real and synthetic errors are similar, and thus the synthetic datasets retain the error structure of the real data. The treatments described in section 2.2.3.1 (above) for applying a minimum error threshold, downweighting “weak” variables, and weighting m/z’s related to m/z 44 were also applied to the error estimates for the synthetic datasets. No error propagation for box smoothing is applied to the synthetic data because these data are not smoothed. The SNR for the m/z’s in the synthetic datasets are therefore higher by a factor of ~sqrt(3), and there are more “weak” variables (84 m/z’s total) than in the real data. The weak m/z’s in the synthetic datasets include the same m/z’s as in the real dataset, as well as m/z’s 150, 185 (2-factor case only), 216, 227, 239 (3-factor case only), and 250-253.

Two-Factor Synthetic Dataset

A two-factor synthetic base case was created using the HOA and OOA MS and TS as determined by Zhang et al. (c; Zhang et al., 2005a) for the Pittsburgh dataset. Difference spectra may contain negative values for very small signals, akin to below-detection limit values in other datasets. Zhang et al. (2005a) allowed their method to fit these small negative values and the resulting factors include small negative numbers. Negative values in the Zhang solution were
converted to their absolute value before creating the $F_{\text{input}}$ and $G_{\text{input}}$ matrices, so that the input has only positive numbers. The resultant increase in signal is much smaller than the residual from the Zhang factorization and does not affect the results of the PMF factorization.

**Three-Factor Synthetic Datasets**

A three-factor synthetic base case was created from the three-factor PMF solution with FPEAK = 0 of the real Pittsburgh dataset (described below). All factor elements were positive for this solution, so no treatment of negative values was necessary.

Variations on the three-factor base case were made to explore the ability of PMF to retrieve factors which have a small fraction of the total mass. Three-factor synthetic cases were created by replacing the mass spectrum of the smallest factor in the previous three-factor synthetic case with reference mass spectra (see Sect. 2.2.4.1) of fulvic acid (FA) (Alfarra, 2004), biomass burning organic aerosol (BBOA, Palmetto leaf smoke from the Fire Lab at Montana Experiment (FLAME-1) in June 2006), or fresh chamber SOA (methylene cyclohexane + O$_3$, (Bahreini et al., 2005)), each of which has a different correlation to the other MS in the input. Variations on this case were made in which the average mass of this factor was decreased (cases with 11.4%, 5.7%, 2.9%, 1.4%, and 0.7% average mass fraction) and used in a new $G_{\text{input}}$ to create a new $X_{\text{input}}$.

**2.2.4 Statistical Comparisons of Mass Spectra**

**2.2.4.1 Reference Spectra**

An AMS Spectral Database (Ulbrich et al., 2007) has been created to collect published AMS spectra for public use. AMS spectra are similar, but not identical to, spectra from standard electron-impact databases such as the NIST Chemistry WebBook (http://webbook.nist.gov/chemistry/ (Stein et al., 2001)), primarily because of thermal
decomposition of molecules during vaporization in the AMS and additional thermal energy of the ions leading to increased fragmentation (Alfarra, 2004; Dzepina et al., 2007). Different AMS instruments operating with the same vaporizer temperature produce similar spectra (Alfarra, 2004; Dzepina et al., 2007). These reference spectra are used to aid the identification of spectra in PMF factors. The database contains spectra for several categories of aerosol, including ambient aerosol, direct measurements from sources (e.g., vehicles, biomass burning), laboratory-generated aerosol of chemical standards, laboratory SOA, laboratory heterogeneously oxidized particles, other laboratory-generated aerosol, and spectra derived from mathematical deconvolutions of ambient OA. Organic aerosol spectra in the database span a range of representative hydrocarbon-like (e.g., diesel bus exhaust, fuel, and lubricating oil) and oxygenated (e.g., various chamber SOA, oxalic acid, and fulvic acid) OA.

2.2.4.2  Statistics of Correlation

Throughout this work we report “uncentered” correlations between MS and TS as a qualitative metric to support factor identification and compare factors amongst different PMF solutions (Paatero, 2008a). The uncentered correlation coefficient (UC) reports the cosine of the angle between a pair of MS or TS as vectors, such that

$$UC = \cos \theta = \frac{x \cdot y}{\|x\| \|y\|}$$  \hspace{1cm} (2.11)

where $x$ and $y$ denote a pair of MS or TS as vectors. The uncentered correlation is very similar to the well-known Pearson R for mass spectra, and quite correlated with Pearson’s R for time series (when computed with a large number of different MS and TS; see Fig. D.1). Correlations between MS are complicated because the signal values span several orders of magnitude and a few high intensity masses (generally all at $m/z \leq 44$) can dominate the correlation (Hemann et
al., 2008). For correlations between factor and reference MS, we also report UC for m/z > 44 to remove this bias (Alfarra et al., 2006; Alfarra et al., 2007; Lanz et al., 2007). These two statistics represent one way to numerically match factor profiles to reference profiles for AMS datasets and improve the source identification process (as suggested by Reff et al., 2007). A rank-correlation method, Spearman R, in which correlations are made using the rank order of values (highest = n, lowest = 1) instead of the actual data values (Press et al., 2007b), was also considered as well as several variations on it (e.g., removing ions with low signal in both spectra before calculation), but this often gave too much weight to small signals and otherwise did not aid interpretation beyond that provided by UC for m/z > 44. Correlations are presented in the text as $UC_{HOA in, HOA out}^{TS}$, $UC_{HOA in, HOA out}^{MS}$, and $UC_{HOA in, HOA out}^{MS, m/z > 44}$, where the superscript describes whether MS or TS are being compared (using only m/z’s > 44 when specified) and the subscript describes what data are being compared (here, the input HOA and output HOA for a synthetic dataset). Correlations are calculated using only the points common to both vectors being correlated; e.g., MS from the AMS Spectral Database may have 300 m/z while factor MS have only the 268 m/z’s that were retained, thus missing m/z values are omitted from the vectors before calculating the correlations; TS from different instruments may be missing different periods of data, thus only the points when both instruments report data are included.

2.3 Results

2.3.1 Real Pittsburgh Data

We explored the effect of data pretreatment (Zhang et al., 2005a), downweighting of “weak” variables by a factor of 2 (Paatero and Hopke, 2003), and use of the robust mode in PMF. Differences in the factor MS and TS were minor in all cases for this dataset. Comparisons
of the 3-factor solutions from the robust and non-robust modes are shown in Fig. D.2. We note, however, that this dataset has good SNR, that pretreatment aids the analysis by removing spikes whose cause is understood (poor sampling statistics of high-mass, low-number particles mainly in HOA \( m/z \)’s (Zhang et al., 2005a)), and that these techniques can make a bigger difference for a dataset with much lower signal-to-noise (Canagaratna et al., 2006). We report results for the case with pretreatment, with downweighting, and in the robust mode in order to capture the broad characteristics of the dataset. Throughout this section, a “case” refers to an input dataset and a “solution” refers to PMF2 results.

2.3.1.1 Solutions as a Function of Number of Factors

Q-values and maximum value of RotMat for the real Pittsburgh dataset for solutions up to 7 factors are shown in Fig. 2.3a and mass fractions of these solutions are shown in Fig. 2.4a, all

![Figure 2.3](image_url)

**Figure 2.3.** Values of \( Q/Q_{exp} \) and the maximum value of RotMat for a) the real Pittsburgh case, b) the two-factor synthetic case, and c) the three-factor synthetic case. d) Percent variance explained by factors from SVD analysis of the Pittsburgh real data matrix and the residual matrices from the Pittsburgh real case solution with 3 factors and the synthetic 3-factor base case solution with 3 factors.
Figure 2.4. Mass fraction of each factor in solutions from 2 to 7 factors for a) the real Pittsburgh case, b) the two-component synthetic dataset, and c) the three-component synthetic dataset.

for FPEAK = 0. There is a large drop in the Q/Q_{exp} value from one to two factors, and Q/Q_{exp} is 1.9 at 2 factors. Additional factors continue to reduce Q/Q_{exp} toward 1, but no strong change in slope is observed (largest steps are 9% from 2-3 factors and 4-5 factors). With seven factors, Q/Q_{exp} is 1.3. The Q criterion clearly implies that at least two factors are necessary to explain the data, but there is no strong indication for choosing another solution. Max(RotMat) has a distinct maximum at 2 factors and much smaller values for larger numbers of factors. There is a local minimum at 3 factors and another at 7 factors (confirmed by solutions with >7 factors).
Based on the trends of Q/Q_{exp} and the max(RotMat) from Lee et al. (1999), solutions with 2 or more factors would be suitable. SVD analysis of the data matrix shows that 3 factors are required to explain at least 99% of the variance (Fig. 2.3d). We proceed to analyze each solution stepwise and attempt to interpret them based on correlations with reference MS and tracer TS and use this interpretability as a guide for choosing the number of factors.

2-factor Solution

The MS and TS of the two-factor solution are shown in Fig. D.3a. These two factors reproduce the MS and TS found by Zhang et al. (2005a) for this dataset using their original 2-component CPCA method \((UC^\text{MS}_{\text{HOA}, \text{Zhang}, \text{HOA PMF}} = 0.98, UC^\text{MS}_{\text{OOA}, \text{Zhang}, \text{OOA PMF}} > 0.99)\). The OOA factor has 12% \(m/z\ 44\) and the HOA factor has 3% \(m/z\ 44\) at FPEAK = 0. All interpretations of the factors made by Zhang et al. (2005a; c) hold for these factors.

3-factor Solution

The MS and TS of the three factor solution at FPEAK = 0 are shown in Fig. 2.5 (correlations between selected PMF and tracer TS are shown in Table D.1). The three-factor solution has HOA and OOA factors very similar to the Zhang et al. HOA and OOA \((UC^\text{MS}_{\text{HOA}, \text{Zhang}, \text{HOA PMF}} = 0.97, UC^\text{TS}_{\text{HOA}, \text{Zhang}, \text{HOA PMF}} > 0.99, UC^\text{TS}_{\text{OOA}, \text{Zhang}, \text{OOA PMF}} = 0.98)\) that correlate well with primary combustion tracers \((UC^\text{TS}_{\text{CO}, \text{HOA}} = 0.93, UC^\text{TS}_{\text{NOx}, \text{HOA}} = 0.95)\) and AMS sulfate \((UC^\text{TS}_{\text{Sulfate}, \text{OOA}} = 0.95)\), respectively. Note that HOA likely encompasses both gasoline and diesel engine emissions, plus other sources of reduced aerosols such as meat cooking.
Figure 2.5. Factors from the three-component PMF solution of the real Pittsburgh dataset for FPEAK = 0. a) Mass spectra of the three components. The fraction of the signal above \textit{m/z} 100 is 3.4%, 24.3%, and 9.7% for OOA-1, OOA-2, and HOA, respectively. b) Time series of the three components and tracers.

(Mohr et al., 2008). PMF analysis of molecular markers results in a similar phenomenon in which the composition of gasoline and diesel emissions are too similar and a factor representing the sum is often retrieved (Brinkman et al., 2006).

The third factor represents 10% of the OA mass and has a MS with strong correlation with several primary, secondary, and biomass burning OA spectra from the AMS Spectral Database (Fig. 2.6a, D.4a, Table D.2). We identify this spectrum as a second type of OOA, OOA-2, because of the strong presence of \textit{m/z} 44 (high resolution aerosol mass spectrometer
Figure 2.6. UC$^{\text{MS}}$ between representative spectra from the AMS Mass Spectral Database (http://cires.colorado.edu/jimenez-group/AMSsd) and a) the third factor mass spectrum from the 3-factor PMF solution of the real Pittsburgh dataset, b) the fourth factor mass spectrum from the 4-factor PMF solution of the real Pittsburgh dataset, and c) the “mixed” factor mass spectrum from the 3-factor PMF solution of 2-factor base case. Values are given in Table D.1. Superscripts denote the source of the reference spectra as follows: (a) Zhang et al., 2005a; b) Canagaratna et al., 2004; c) Alfarra et al., 2004; d) Alfarra, 2004; e) Bahreini et al., 2005; f) Sage et al., 2007; g) I.M. Ulbrich, J. Kroll, J.A. Huffman, T. Onash, A. Trimborn, J.L. Jimenez, unpublished spectra, FLAME-I, Missoula, MT, 2006; h) Schneider et al., 2006).
data of ambient aerosols confirm that this is most likely CO$_2$\(^+\) (DeCarlo et al., 2006; Huffman et al., 2009)), and the correlation with OOA/SOA spectra. The OOA that accounts for most of the mass is very similar to that identified by Zhang et al. (2005a) and Lanz et al. (2007) is now termed OOA-1, following the nomenclature of Lanz et al. (2007). The OOA-2 spectrum lies 23 degrees out of the HOA/OOA-1 plane (calculation described in Appendix D), is clearly not a linear combination of the HOA and OOA-1 spectra, and is unlikely to arise due to noise. The lack of significant \(m/z\)’s 60 and 73 strongly suggests that this OOA-2 does not arise from a biomass burning source (Alfarra et al., 2007; Schneider et al., 2006). The OOA-2 time series correlates well with ammonium nitrate and ammonium chloride from the AMS, two secondary inorganic species which were not included in the PMF analysis (\(U_{\text{Ammonium Nitrate, OOA-2}}^{\text{TS}} = 0.79, U_{\text{Ammonium Chloride, OOA-2}}^{\text{TS}} = 0.82\); diurnal cycles are shown in Fig. D.5). Note that we can confirm that nitrate and chloride signals from the AMS are indeed dominated by the inorganic ammonium salts, not fragments of organic species, based on the ammonium balance (Zhang et al., 2005b; Zhang et al., 2007b). The OOA-2 factor is less-oxygenated than OOA-1 and more oxygenated than HOA (\(m/z\) 44 of OOA-1 is 12.5%, of OOA-2 is 6%, and of HOA is 2.5%). Since both nitrate and chloride show a semivolatile behavior in Pittsburgh (Zhang et al., 2005b), these correlations imply that OOA-2 is also semivolatile. Most likely OOA-2 corresponds to less oxidized, semivolatile SOA, while OOA-1 likely represents a more aged SOA that is much less volatile. Direct volatility measurements with a thermal-denuder AMS combination indeed show that in Mexico City and Riverside, CA, the less oxygenated OOA-2 component is more volatile than the OOA-1 component (Huffman et al., 2009). A similar OOA-2 factor with a less oxidized spectrum and a high correlation with nitrate was reported by Lanz et al. (2007) for their dataset in Zurich in summer of 2005, though the ratios of OOA-2 to nitrate differ (~1 in the present
work, ~2 in Zurich). These authors also interpreted OOA-2 as fresh SOA. No evidence is available to support the identification of OOA-1 or OOA-2 as either “anthropogenic” or “biogenic” in origin.

4-Factor Solution

The TS and MS for the 4-factor solution are shown in Fig. D.3b. The four-factor solution has clear HOA, OOA-1, and OOA-2 factors with high similarity to those in the 3-factor case (\(UC_{\text{TS}}^{\text{factor HOA, factor OOA-1, factor OOA-2}} = 0.99\); \(UC_{\text{MS}}^{\text{factor HOA, factor OOA-1, factor OOA-2}} = 0.99\)) which are interpreted as in the 3-factor case. The fourth factor correlates well with sulfate (\(UC_{\text{TS}}^{\text{amine factor Unnamed, sulfate}} = 0.84\)) and has a similar mass fraction as OOA-1 (39% and 24%, respectively). The MS of the fourth factor has \(UC_{\text{MS}}^{\text{database spectra, factor Unnamed}} > 0.8\) with database mass spectra of the Zhang Pittsburgh OOA, three types of SOA, and four types of biomass burning (Fig. 2.6b, D.4b, Table D.2). However, strong independent evidence (such as a strong tracer correlation) for a distinctive, separate component is not present. This additional component shows many (but not all) of the signs of the “component splitting” of the solutions of the synthetic data cases discussed below (Sect. 2.3.2.1). In addition, the analysis of the residuals (below) does not suggest that adding components after the first 3 fits more of the data. Thus in the absence of any supporting evidence, we concluded that this component represents an artificial “splitting” of the solution (calling this factor OOA-1a) and that keeping this component would be an overinterpretation of the PMF results. In particular we warn about trying to interpret e.g. one of the OOA-1’s as “biogenic” and the other as “anthropogenic” or similar splits, in the absence of strong evidence to support these assignments.
Five and more factor solutions

The five-factor solution (Fig. D.3c) has four factors that are similar (OOA-1, OOA-1a, OOA-2, HOA) ($U_{4\text{-factor,5\text{-factor}}}$ > 0.96, $U_{4\text{-factor}}^{\text{MS}}$ > 0.90) to the factors in the 4-factor solution. The fifth factor (HOA_a) is similar to the HOA factor in this solution ($U_{5\text{-factor HOA,5\text{-factor HOA}_a}}^{\text{MS}}$ = 0.85, $U_{5\text{-factor HOA,5\text{-factor HOA}_a}}^{\text{TS}}$ = 0.68) and has $U_{\text{5\text{-factor HOA} Database Spectra,5\text{-factor HOA}_a}}^{\text{MS}}$ > 0.8 with five types of SOA and eight types of BBOA, but there is no strong correlation with any available tracer. The HOA and HOA_a factors have 23% and 15% of the mass, respectively. As before, we conclude that this “splitting” of the HOA is most likely a mathematical artifact and not a real component.

Interpretation of factors in the six- and seven-factor solutions becomes more complex and no independent information from tracer correlations exists to substantiate the interpretation of these factors. These factors likely arise due to splitting of the real factors, likely triggered by variations in the spectra of the real components (discussed below).

Uncertainty of the Solutions of Real Data

In order to explore the possibility of multiple local minima in the solutions of the dataset and qualitatively assess variability in the factors, trials with 64 multiple starts were calculated for the real Pittsburgh case with solutions up to 6 factors. Local minima can be identified by solutions with different $Q/Q_{\text{exp}}$ values, but this is not a sufficient criterion as it could be possible for two local minima to have similar $Q/Q_{\text{exp}}$-values with different factors; therefore similarity of the factor MS and TS is also considered as a criterion for determining local minima. In the solutions with 2- to 6- factors, no local solutions were observed. The 3-factor solutions show the
greatest variation in $Q/Q_{\text{exp}}$ values, which however increase by only $2 \times 10^{-4}$ $Q/Q_{\text{exp}}$ units above the minimum. There are two modes of the solutions in this small range, defined by the ratio of $m/z$ 43: $m/z$ 44 in the MS of the OOA-2 factor, while the MS of the OOA-1 factor varies little and the MS of the HOA factor is virtually identical in all solutions. The TS of all of these solutions are virtually identical (the overlaid TS and MS of all 64 solutions for the 3-5 factor solutions are shown in Fig. D.6).

Quantitative assessment of the uncertainty of the factors is made by 100 bootstrapping runs (Norris et al., 2008) of which 95 reproduce all three factors. The average factor MS and TS along with $1\sigma$ variation bars for each point are shown in Fig. D.7. These results show that the uncertainty in the TS of the three components is small, as it is for the MS of HOA and OOA-1. The MS of OOA-2 shows some uncertainty, which is nevertheless small compared to the general structure of the spectrum. Thus we conclude that the results reported here are robust and that their statistical uncertainties are small.

**Residuals of the PMF solutions**

Figure 2.7 shows the $Q/Q_{\text{exp}}$ values for each point in time and Fig. D.8 shows the total residuals ($\Sigma$ residual), total absolute residuals ($\Sigma \left|\text{residual}\right|$), and normalized absolute residuals ($\Sigma \left|\text{residual}\right| / \Sigma \text{signal}$) for the 3- through 6-factor solutions for $F\text{PEAK} = 0$. Note that $Q_{\text{exp}}$ for a time sample equals the number of $m/z$'s in the MS (268). Figure 2.8 shows a summary distribution of the scaled residuals for all $m/z$ from the real Pittsburgh data and Fig. D.9 shows the distribution of the scaled residuals for selected $m/z$'s; minimization of total $Q$ (squared scaled residuals) while meeting non-negativity constraints drives the solutions of the PMF algorithm. The contributions to both $Q/Q_{\text{exp}}$ and absolute residual, even after fitting 3 to 6 factors, have
considerable structure (above a background level) that is very similar to the OOA-2 time series from the three-factor solution. The UC values for the OOA-2 contribution from the three-factor solution are as follows: $U_C^{TS}$ for the OOA-2 contribution is 0.70, 0.75, and 0.78 for the four- and five-factor solutions, respectively. Adding more factors results in only minor changes in the TS of Q/Q$_{exp}$ contributions and residual, implying that the same data variation fit by the lower order solution is being refit with more factors. In fact, the decrease in the TS of Q/Q$_{exp}$ (improvements in the fit) in solutions with 3 to 6 factors do not occur during periods of high HOA or OOA-2, and only occasionally during periods of high OOA-1 (Fig. D.10). The highest
Figure 2.8. a) Distributions of scaled residuals for each m/z for the 3-factor solution of the real Pittsburgh case. Black markers represent medians. Green boxes span the 25th and 75th percentiles; blue whiskers attached to the boxes extend to 10th and 90th percentiles. Grey “floating” whiskers connect 2nd to 5th percentiles and 95th to 98th percentiles. b) Expansion of a) showing the scaled residuals from -4 to 4.

spike in the residual TS, a short-lived event on the evening of 9/14/02, is likely due to a specific HOA plume (e.g., a specific combustion source) whose spectrum is similar to but has some differences from the main HOA factor during the study and shows variation in m/z peaks with higher contribution to HOA than OOA-2. SVD of the unscaled residual matrix after fitting 3 factors (Fig. 2.3d) shows that with even 12 more factors, less than 95% of the remaining variance can be explained (150 factors would be needed to explain 95% of the variance in the matrix of
scaled residuals that was not downweighted for weak m/z’s or those proportionally related to m/z 44). The residual at specific m/z’s during periods of high OOA-2 and high Q/Qexp changes for many significant OOA-2 m/z’s in modest amounts, fairly continuously, over periods of 10-20 minutes. This is likely caused by variations in the true OOA-2 spectrum (which could occur, e.g., during condensation or evaporation of SVOCs) that cannot be represented by the constant-MS factor, nor are constant enough to become their own factor. These behaviors imply that three factors have explained as much of the data as is possible with a bilinear model with constant spectra.

2.3.1.2 Rotations

The three factor solution, which is the most interpretable as discussed above, is tested for its rotational ambiguity. The FPEAK range required for the Q/Qexp = 10% criterion in the real data is -4.2 to +4.4. Solutions with FPEAKs between -1.6 and +1.0 give an increase of 1% over Q/Qexp at FPEAK = 0 (Fig. 2.9). MS and TS spanning this range of solutions are shown in Fig.

**Figure 2.9.** Q/Qexp vs. FPEAK for the real Pittsburgh case with 3 factors, the 2-factor synthetic base case with 2 factors, and the 3-factor synthetic base case with 3 factors.
2.10. Note that changing FPEAK changes both the MS and TS simultaneously. Overall, the effect of positive FPEAK is to create more near-zero values in the MS and decrease the number of near-zero values in the TS. The effect of negative FPEAK is to create more near-zero values in the TS and decrease the number of near-zero values in the MS. Note for example that the TS of the FPEAK = -1.6 solution have periods of zeros that do not correlate with any interpretable events, likely indicating that this solution represents rotation beyond the range that gives useful insight for this dataset. Changes in TS occur more in some periods than others. Mass
concentration of all factors remain fairly constant at all FPEAKs during periods in which at least one factor has a mass concentration near zero, but periods in which all factors have non-zero mass concentrations show more variation as FPEAK is changed. This is most dramatic for the OOA-2 events on 9/13/2002 and 9/14/2002, in which negative FPEAKs give more mass to the OOA-2 factor and less mass to the OOA-1 factor compared to solutions with FPEAK $\geq 0$. These differences represent one way of characterizing the uncertainty of the PMF solutions, since the Q/Q_{exp} values change little between them and all the TS and MS appear physically plausible. The solutions from multiple FPEAKS (Fig. 2.10) show a greater range in MS than the bootstrapping 1-$\sigma$ variation bars, while the TS show a similar range to the bootstrapping 1-$\sigma$ variation bars (Fig. D.7).

The MS change with FPEAK is most dramatic in the OOA-2 MS, while the OOA-1 MS changes very little with FPEAK. This is not surprising since OOA-2 accounts for a low fraction of the total signal and thus its spectrum can change more without causing large increases in the residuals. At large negative values of FPEAK, the OOA-2 factor strongly resembles the HOA factor ($U_{OOA-2,\text{HOA at FPEAK = -1.6}}^{\text{MS}} = 0.98$). The ratio of m/z 43 to m/z 44 in OOA-2 decreases from 2.1 to 1.1 and 0.55 as FPEAK increases from -1.6 to 0 and +1.0, respectively. A sharp decrease in the fraction of signal attributed to m/z 55 relative to a small decrease in m/z 57 (ratios of 0.88, 0.50, and 0 at FPEAKs -1.6, 0, and +1.0, respectively) gives m/z 57 an unusually high fraction of the signal in OOA-2 at large positive FPEAKs. Positive FPEAK values also reduce the fraction of m/z 44 (mainly CO$_2^+$) attributed to the HOA MS.

Diagnostic graphs of the correlations between each of the three factors from the real case, and how these correlations change with FPEAK, are shown in Fig. 2.11a (Fig. D.11a). The OOA-2 factor is highly correlated in MS with both the HOA and OOA-1 factors at FPEAK = 0.
Figure 2.11. Correlations for PMF factors to each other as they change with FPEAK for a) the real Pittsburgh case and b) the three-factor synthetic base case. The red labels denote the correlations of the input factors. Traces are colored by FPEAK and numbers denote the FPEAK of each solution. Black dots in b) indicate the correlation of the factors in the input. c) Correlation of the PMF factors to the input factors for the three-factor synthetic case. d) Expansion of c).

\( UC_{\text{OOA-2,HOA}}^{\text{MS}} = 0.81 \), \( UC_{\text{OOA-2,OOA-1}}^{\text{MS}} = 0.84 \), respectively, but has a more correlated time series with HOA than with OOA-1 \( UC_{\text{OOA-2,HOA}}^{\text{TS}} = 0.84 \), \( UC_{\text{OOA-2,OOA-1}}^{\text{TS}} = 0.55 \). As expected, the behavior of positive values of FPEAK is to mix the time series, making them more correlated and the mass spectra less correlated, moving these solutions toward the upper left of the graph.

Negative values of FPEAK do the opposite. Note that this type of graph is very sensitive to
small changes in the solution TS and MS. We find no evidence that solutions with FPEAKs
away from zero are preferable.

2.3.2 Synthetic AMS Data

It is clear from the analysis of this ambient dataset that there is significant ambiguity in
the “correct” choice of the number of factors and rotations of the solutions. In this section we
seek to evaluate the behavior of PMF2 and the appearance of its solutions for AMS synthetic
datasets for which the true solution is known and for which well-specified and realistic noise is
the only additional content in the data matrix. This type of analysis will help in the interpretation
of the solutions from real cases.

2.3.2.1 Solutions of Synthetic Data Base Cases

Solutions as a Function of Number of Factors

As AMS organic datasets have a specific structure with strong interrelation across m/z’s
and auto- and cross-correlation time scales of the order of hours to days for most components
and datasets, it is of interest to examine what PMF2 reports when too few or too many factors are
requested by the user when the correct number of factors is known.

Q/Q_{exp}-values and maximum value of RotMat for the two- (and 3-) factor synthetic base
case for solutions up to 7 factors are shown in Fig. 2.3b (2.3c) for FPEAK = 0. Q/Q_{exp} is \gg 1
for the one-factor solution and is \approx 1 for all solutions of 2 (3) or more factors. This is the
expected behavior; one factor should not fully explain the data and a high residual is reasonable.
With the addition of a second (third) factor, most of the dataset is explained and the residual is
on the order of the noise. As the simulated noise has a large number of degrees of freedom,
solutions with more than two (three) factors are not able to explain more of the data and $Q/Q_{\text{exp}}$ is approximately constant. $\text{Max}(\text{RotMat})$ does not show a clear trend with number of factors. The maximum is at three (two) factors and a local minimum occurs at six (four) factors. No criterion for choosing the correct number of factors from $\text{max}(\text{RotMat})$ can be determined from these cases.

Mass spectra and time series for the 2- to 5- factor solutions of the 2- and 3-factor synthetic base cases are shown in Figs. D.12 and D.13. Mass fractions of the 2- to 7-factor solutions of the 2- and 3-factor synthetic base cases for $\text{FPEAK} = 0$ are shown in Figs. 2.4b and 2.4c. The correct number of factors (2 or 3) reproduces the input factors ($UC_{\text{MS}}^{\text{MS}}$, $UC_{\text{TS}}^{\text{TS}}$, $UC_{\text{MS}}^{\text{MS}}$, $UC_{\text{TS}}^{\text{TS}}$) very well. When more factors are calculated, factors highly similar to the input factors are retained and the additional factors have strong resemblance to one of these factors. For example, the 4-factor solution of the 3-factor base case has HOA and OOA-2 factors highly similar to the input for these factors ($UC_{\text{MS}}^{\text{MS}}$, $UC_{\text{TS}}^{\text{TS}}$, $UC_{\text{MS}}^{\text{MS}}$, $UC_{\text{TS}}^{\text{TS}}$) but two OOA-1 factors appear ($UC_{\text{MS}}^{\text{MS}}$, $UC_{\text{TS}}^{\text{TS}}$, $UC_{\text{MS}}^{\text{MS}}$, $UC_{\text{TS}}^{\text{TS}}$) with 42 and 20% of the fit mass, respectively. The OOA-1 factor has been “split” into two. This behavior continues with more factors in both base cases. Split factors have very similar MS and TS to each other and to the same factor in the $p$-factor solution with approximately equal splitting of the mass between the like factors.

The 3-factor solution of the 2-factor base case shows another interesting behavior (Fig. D.12b). One factor is very similar to the HOA input factor ($UC_{\text{MS}}^{\text{MS}}$, $UC_{\text{TS}}^{\text{TS}}$, $UC_{\text{MS}}^{\text{MS}}$)
UC_{\text{HOA Input, 3--factor HOA}}^{\text{TS}} > 0.99) and one factor very similar to the OOA input factor (UC_{\text{OOA Input, 3--factor OOA}}^{\text{TS}} > 0.99). The third factor’s spectrum includes high-signal peaks from both the HOA and OOA spectra. It is highly correlated with both the input HOA and OOA spectra (UC_{\text{HOA Input, 3--factor Mixed}}^{\text{MS}} = 0.78, UC_{\text{OOA Input, 3--factor Mixed}}^{\text{MS}} = 0.91, whereas the correlation between the input HOA and OOA spectra (UC_{\text{HOA Input, OOA Input}}^{\text{MS}}) is only 0.43). This third factor lies 3.1° out of the plane of the HOA and OOA spectra. A true linear combination would be in the HOA/OOA plane (0°); in two-factor synthetic cases without added Poisson and electronic noise, the third factor is a true linear combination of the HOA and OOA spectra and lies in the plane of the HOA and OOA factors. Thus it is the presence of the noise that allows the third factor to be slightly outside the 2-component plane. This factor represents a "mix" of the true HOA and OOA spectra, especially because of the increased fraction of m/z 44 compared to the input HOA (<1% in the HOA input, 7% in the mixed factor). We therefore choose to call it a “mixed” HOA/OOA factor. This “mixed” factor, however, has a time series that is similar to the HOA time series (UC_{\text{HOA Input, 3--factor Mixed}}^{\text{TS}} = 0.97) and it takes 11% of the total mass, taking more from HOA than OOA mass. Because we created a 2-factor input, we know that this is not a true factor. Nevertheless, the profile of this “mixed” factor is very similar (UC_{\text{Mixed Factor, Database Spectra}}^{\text{MS}} > 0.9) to multiple mass spectra from the AMS Spectral Database as shown in Fig. 2.6c (Fig. D.4c, Table D.2), including SOA from chamber reactions (e.g., β-caryophyllene + O₃, α-terpinene + O₃), and particles from combustion of biomass or biomass components (e.g., ceanothus, juniper). In fact, this “mixed” spectrum has higher correlation with many of the SOA and BBOA spectra than the OOA-2 spectrum from the real Pittsburgh case (Fig. 2.6a). No mixed factors are observed in the 3-factor base case, indicating that the tendency
of PMF to “mix” or “split” factors in its solutions (when too many factors are requested) is not easily predicted and can shift with relatively small changes in the input matrix (such as the addition of an OOA-2 input factor with 10% of the mass here).

**Uncertainty of the Solutions of Synthetic Data**

In order to explore the possibility of multiple local minima in the solutions and the repeatability of the splitting behavior of the synthetic datasets, as well as to qualitatively assess variability in the factors, trials with 64 multiple starts were calculated for the 2-factor and 3-factor synthetic cases with solutions up to 6 factors. Similar to the real case, the range of $Q/Q_{exp}$ values across the seeds was quite small, with the greatest difference in the 2- (3)-factor case of $2 \times 10^{-4}$ ($1 \times 10^{-6}$) $Q/Q_{exp}$ units above the minimum. No local minima were identified in the 3-factor synthetic cases. In the 2-factor synthetic cases, local minima are identified in solutions with 3 or more factors. The solutions fall into groups in which one factor represents the input OOA, one factor the input HOA, and the other factors shows various degrees of mixing of the HOA and OOA factors, as was described in Eqs. (2.5, 2.6). The three-factor solution of the 2-factor synthetic case shown in Fig. D.12b is representative of the most common type of solution identified. Bootstrapping analysis of the synthetic cases with the correct number of factors results in 1-σ variation bars of the same order or smaller than in the real case.

**Rotations of the 3-Factor Synthetic Base Case Solution**

In this section we examine the range of “rotatability” of the solutions. The FPEAK values required for the $Q/Q_{exp}$ increase from its minimum by e.g. 10% in synthetic data can be quite large; $\text{FPEAK} = \pm 3.0$ for the 2-factor case gives an increase in $Q/Q_{exp}$ of only about 4%, but an FPEAK range of -1.0 to +1.0 in the 3-factor case gives an increase in $Q/Q_{exp}$ of 1% (Fig.
2.9). MS and TS spanning this range of solutions are shown in Fig. D.14. Changes in the MS and TS are similar to those in the real Pittsburgh case. In the TS, mass concentrations during periods in which all factors have non-zero mass concentration show more variation as FPEAK is changed and remain approximately constant in periods in which at least one factor has near-zero mass concentration. The MS change with FPEAK is most dramatic in the OOA-2 MS, while the HOA and OOA-1 MS change very little with FPEAK. The changes of the OOA-2 MS with FPEAK are similar to those of the real Pittsburgh case, with a strong resemblance to the HOA factor at negative FPEAK values \( UC_{\text{MS, OOA--2 at FPEAK<--1}} > 0.95 \) and change of its fraction of \( m/z \) 44 with increasing FPEAK values, from 6% at FPEAK 0 to 8% at FPEAK 1.0.

A graphical diagnostic of the correlations of each of the three factors versus each other and how they change with FPEAK is shown in Fig. 2.11b (Fig. D.11b). The input factors are reproduced best near 0 FPEAK values (Figs. 2.11c,d, D.11c,d). The effect of FPEAK on the correlation of the factors is extremely similar to that in the real Pittsburgh case, with positive values of FPEAK making them more correlated and the mass spectra less correlated and negative values of FPEAK doing the opposite, making the time series less correlated and the mass spectra more correlated. The length of the segment between adjacent FPEAK solutions indicates the relative amount of change between the factors.

**Residuals in the 3-Factor Synthetic Base Case**

The Q/Q_{exp} contribution as a function of time for the 1-, 2-, and 3-factor solutions of the 3-factor base case are shown in Fig. 2.7c. The 1- and 2-factor solutions have distinct structure which indicates that all of the variation in the data has not been fit. In the three-factor solution, which reproduces the three input factors very well \( UC_{\text{MS, Input,3,factor Output}} ^{\text{MS, Input,3,factor Output}} > 0.99 \)
for all three factors), the Q/Q_{exp} contribution has no visible structure. SVD of the unscaled residual matrix (Fig. 2.3d) shows that even with 15 more factors, only about 75% of the remaining variance can be explained. These observations indicate that only random noise remains.

### 2.3.2.2 Separation of Correlated Factors

Another important question when analyzing PMF results is: how similar (correlated) can the factors be (in MS, TS, or both) and still be correctly retrieved by PMF? In one extreme case, two different sources that are completely correlated in time (UC_{TS} \sim 1) cannot be separated by PMF or any other factorization method, as there is no information for their separation. Even if their MS are very different, only the sum of the two sources can be retrieved. Similarly, two sources with very highly correlated MS (UC_{MS} \sim 1) cannot be retrieved separately, irrespective of how different their time series are; again, only their sum can be retrieved. If the correlation between the input factors is plotted in a graph with UC_{MS} as the x-coordinate and UC_{TS} as the y-coordinate (as in Fig. 2.11a,b), the “irretrievable” regions include the edges of the plot near $UC_{TS}^{\text{Factor } i, \text{Factor } j} = 1$ or $UC_{MS}^{\text{Factor } i, \text{Factor } j} = 1$. It would be of interest to evaluate how far two factors have to be from those edges to be retrievable with some accuracy by PMF, and how the rotational uncertainty changes as the factors are closer or farther from those edges. Likely this “distance” from the edges will not be an absolute criterion, but will depend on the fraction of the mass accounted by each factor, i.e. a small factor may be harder to retrieve in this situation since its variance is smaller compared to noise and potential variations in the spectra of the larger components. This is a complex mathematical problem, complicated by the need to understand and maintain the number and placement of “near-zero” values in the factor MS and TS while changing the correlations between the input factors (Paatero, 2008a; Ulbrich et al., 2008). In the
present work, only the retrieval of factors with a given correlation and varying fractions of the mass is explored.

**Separation of Small Factors: Three-factor Synthetic Cases**

The correlations between the factors in the variations on the 3-factor synthetic base case (described above in Sect. 2.2.3) are shown in Fig. 2.12a (Fig. D.15a, Table D.3). Replacing the OOA-2 MS with one of three database spectra (BBOA, SOA, or fulvic acid (FA) MS) gives a broad range of correlations of the third factor with HOA ($U_{\text{MS}}^{\text{HOA,BBOA}} = 0.88$, $U_{\text{MS}}^{\text{HOA,OOA-2}} = 0.81$, $U_{\text{MS}}^{\text{HOA,SOA}} = 0.75$).

**Figure 2.12.** a) Correlations between 3-factor real output and all synthetic input cases. b) Retrieval correlations between PMF and input TS and MS versus mass fraction of different third factors in synthetic cases. Markers denote the resemblance of the factors to the input MS or TS. For the MS, O, S, F, and B refer to OOA-2, SOA, FA, and BBOA, respectively. For the TS, T refers to the input TS. For both cases, H and I refer to HOA and OOA-1, respectively.
$UC_{\text{MS HOA, SOA}} = 0.76$, $UC_{\text{MS HOA, FA}} = 0.39$) and a narrower range of correlations between OOA-1 and the third factor (from $UC_{\text{MS OOA=1, BBOA}} = 0.81$ to $UC_{\text{MS OOA-1, FA}} = 0.89$). We now consider how well PMF can retrieve these factors with average mass fractions between 0.7% and 11.4%.

The ability to retrieve the small factor is assessed by correlation with the input factor (Figs. 2.12b, D.15b). For mass fractions of 11.4%, the small factor was retrieved well at FPEAK = 0 with $UC_{\text{MS OutputFactor, Small Input, Factor}} > 0.99$, $UC_{\text{TS OutputFactor, Small Input, Factor}} > 0.97$. As the average mass fraction of the small factor is decreased, the small factor is not retrieved as accurately in some cases. Both the MS and TS of the small factor can be retrieved with $UC_{\text{MS Small Factor Input, Small Factor Output}} > 0.97$ (except for the SOA case) and $UC_{\text{TS Small Factor Input, Small Factor Output}} > 0.99$ when the small factor has at least 5% of the average mass. As the average mass fraction of the small factor decreases, the third factor may have greater similarity to HOA or OOA-1 in MS or TS than to the input factor. These behaviors have elements similar to the splitting behavior observed in the synthetic base cases when too many factors were chosen. Although the mass spectrum of the third factor input does not change as the mass is reduced, the ability of PMF to retrieve both the mass spectrum and time series decreases as the fraction of the mass decreases. Based on this case study, we conclude that in Q-AMS datasets with only Poisson ion noise and Gaussian electronic noise, factors with a mass fraction of at least 5% are likely to be retrievable with sufficient accuracy for reliable interpretation. This should be a lower limit for real cases in which the noise may have additional structure. However if the small factor has a very different time series from those of the main factors (e.g. if it is a spiky local source or has a distinctive MS) factors with smaller mass fractions may still be reliable retrieved (Huffman et al., 2009).
2.4 Discussion

In the ambient and synthetic data we compared various criteria for determining number of factors. The criterion for choosing the number of factors based on the behavior of $Q/Q_{exp}$ vs. number of factors has the expected behavior for the synthetic datasets but does not lead to a strong conclusion for the real data case (Fig. 2.3). The $Q/Q_{exp}$ of the real dataset has a steep decrease from 1 to 2 factors, but a steady decrease with more factors does not point to a particular choice for number of factors. Our choice of a three-factor solution for the real case is justified by tracer correlations, spectral comparisons and interpretation (e.g., $m/z$ 44 as a surrogate for O:C ratio), and residual analysis, and is a good physical explanation of the data. We conclude that a criterion based on a significant decrease in $Q/Q_{exp}$ as the number of factors is increased is not sufficient to determine the correct number of factors in AMS datasets.

The criteria of the smallest maximum value of RotMat for the best solution proposed by Lee et al. (1999) is not met by any of the datasets. There is no theoretical reason to support the use of the max(Rotmat) criterion to determine the number of factors. The use of the max(Rotmat) criterion appears to be based in the assumption that true data should not have rotational ambiguity, and use of this criterion may favor the wrong number of factors if the data do have significant rotational ambiguity (Paatero, 2008b). In the synthetic datasets, max(RotMat) gives no indication about the correct number of factors. Since the criterion is not met for synthetic datasets, we draw no conclusion from this statistic for real datasets and do not recommend this as a basis for choosing the correct number of factors in AMS data. RotMat values are larger in the real case than for the synthetic case, suggesting additional rotatability in the real case.
The splitting and mixing behavior seen in the synthetic datasets when too many factors were requested presents a clear warning that when evaluating a real dataset, the researcher must be careful about overinterpreting solutions with multiple factors, even though it may be tempting to “identify” additional sources and give them physical-sounding names. The characterized splitting behavior in which each split factor has comparable fractions of the mass of the original factor may be specific to the PMF2 algorithm, as other algorithms such as MCA (Zhang et al., 2007a) tend to split a factor into a dominant factor (in terms of mass) and a small factor. In particular, care must be taken because the spurious factors obtained due to “splitting” or “mixing” behavior have realistic-looking mass spectra and time series and could easily be interpreted by the unsuspecting user as a real factor, though we know that in the synthetic cases these factors must be mathematical artifacts. In cases where mixing behavior occurs, non-existent factors could be retained. Thus, it is critical to use external tracers to confirm the interpretation. In the absence of these tracers, the lower order solution may be the best choice. In our ambient case, for example, factors 3-6 are mathematically good solutions. The fourth factor in the 4-factor solution of the real Pittsburgh dataset exhibits some of the behavior characterized by the splitting observed in the synthetic cases. The TS of this factor is very similar to the TS of the OOA-1 factor (Fig. D.3b) and these two factors share approximately equal portions of the mass (Fig. 2.4a). The MS correlates well with many types of OA measured by the AMS (Fig. 2.6b). While this factor may reflect true variability in OOA-1, we do not have an independent tracer to support this interpretation and prefer the 3-factor solution as the safest choice to avoid overinterpretation.

A key concern in having ambiguity in number of factors is whether the mass fraction of total OOA and HOA (and other major components such as BBOA) determined from a given
PMF analysis depends on the number of factors that is used. The mass fraction of total HOA and OOA varies by approximately ±10% in the real case and in the three-factor synthetic case and could be over- or underestimated depending on the choice of number of factors (Fig. 2.4). This determination is more complex in the two-factor synthetic case and depends strongly on whether the “mixed” factor is attributed to HOA or OOA, or to a new type of source. For solutions with more than 2 factors, both total HOA and total OOA could be underestimated while the mixed factor spuriously accounts for 9-26% of the mass. In the real case, the 2-factor solution slightly underestimates OOA. This behavior is counter to that seen by Lanz et al. (2007), where they observed that a 2-factor solution significantly underestimated OOA in Zurich in the summer. Nemitz et al. (2008) found that the 2-factor solution significantly underestimated OOA by 17% in their case. Thus the tendency for a 2-factor solution to over- or underestimate the OOA/HOA ratio is most likely dependent on the structure of a particular AMS dataset and no general trend is apparent.

In the ambient and synthetic cases we explored the use of FPEAK to explore rotations of the solutions. A ΔQ/Q_{exp} over the minimum Q/Q_{exp} of ~1% (rather than 10%, P.K. Hopke, Pers. Comm. 2007) seems to give a more appropriate range of FPEAKs for this Q-AMS dataset. Positive values of FPEAK mix the TS, making them more correlated and the MS less correlated, while negative values of FPEAK mix the MS, making them more correlated and the TS less correlated (Fig. 2.11). Choice of a particular FPEAK solution is complex. Mathematically, values of zero in the true factors (in MS or TS) help to constrain the rotation, but they must be known a priori to justify a particular FPEAK solution (Paatero et al., 2005; Paatero et al., 2002). We have no such a priori information in the Pittsburgh case to constrain any values in the TS or MS to zero, and so have no mathematical way to choose a particular value of FPEAK. We have
followed the guidance of Paatero et al. (2002) and present a range of solutions for our dataset to describe the degree of rotational ambiguity (Fig. 2.10).

Is it possible, then, to support the choice of any particular rotation? Paatero et al. (2005) suggest use of a graphical method for choosing a rotation such that the source contribution factors (TS) show weak statistical independence near the x- and y-axes when contributions for pairs of factors are plotted in scatter plots. In AMS data, this usually requires less-correlated TS, i.e., negative FPEAKs. Based on our interpretation of the factors as primary OA, fresher SOA, and aged SOA, the concentrations of the factors may not be independent but may be linked by increases and decreases in regional dispersion and photochemistry, and such a rotation may not be warranted. FPEAKs from -1.2 to -1.6 give third factors (OOA-2 in solutions with FPEAKs > -1.2) that exhibit some of the behaviors of splitting (they are similar to HOA, i.e., too close to the right edge of Fig. 2.11a), but this may not be sufficient for rejecting them. The solutions from these FPEAKs create zeros in the TS that do not correspond to interpretable events (Fig. 2.10), and this may be sufficient qualitative justification for rejecting these solutions. No behaviors are observed in the solutions at large positive FPEAKs to support exclusion of these solutions. A possible way to choose a rotation would be to maximize the correlation of the factor TS with external tracers. Though this has no mathematical basis, it could be justified when the researcher is confident of his/her interpretation of the factors. Note, however, that the interdependence of MS and TS during rotations strongly suggests that the change in MS with FPEAK also be considered. Figure 2.13 (Fig. D.16) shows the correlation between each factor MS and TS with selected reference spectra and tracer species, respectively, versus FPEAK. While correlations with TS could support the choice of a positive FPEAK, correlations with the MS suggest that extreme FPEAKs distort the MS and FPEAKs closer to 0 appear more reasonable. Though this
Figure 2.13. Correlation versus FPEAK between PMF factor a) MS and selected reference MS, and b) TS and tracer TS.

is not a very strong criterion for this dataset, we believe that the FPEAK = 0 solution is the best representation of the retrievable factors for this case.

Separation of highly-correlated factors is a potential limitation of this technique with Q-AMS data. HR-ToF-AMS datasets contain more chemical information at each nominal m/z with both oxygenated and non-oxygenated ions (e.g., C$_2$H$_3$O$^+$ and C$_3$H$_7^+$ at m/z 43) that will reduce
the UC_{MS} between true high-resolution components, increasing the retrievability of factors in those datasets (Aiken et al., 2009). Use of diagnostic plots such as those in Fig. 2.11 in Q-AMS and HR-ToF-AMS factorization is encouraged. We feel that great care must be taken when interpreting solutions in which output factors have MS and/or TS with UC > 0.9, especially if they exhibit some characteristics of splitting or mixing behavior.

Studies with 3-factor synthetic cases demonstrate that factors with a small average mass fraction may not be accurately retrieved. Factors with at least 5% of the mass were retrieved well in all cases studied. Factors with smaller mass fractions were often poorly retrieved, with their mass spectra not found and instead a 3^{rd} component appeared due to behavior similar to “splitting” or “mixing” of the 2 dominant factors (Fig. 2.12). The inability to retrieve factors with a smaller fraction of the mass is a likely limitation of this technique with Q-AMS data.

It is important to note that we do not assign the factors of the AMS data to specific sources. For example, the calculated HOA spectrum is likely a linear combination of the HOA sources sampled across the study (encompassing both the range of MS of the sources and their relative mass fractions). For example, separating factors for diesel and gasoline vehicle emissions is very challenging in this analysis because the MS as measured in the AMS are extremely similar and a single HOA MS can represent both sources (and others with similar unit-resolution MS (Mohr et al., 2008). Therefore reported HOAs may vary between studies depending on the specific mix of sources measured during each study. Similarly, the types of OOA reported in each study relate to the distribution of precursors and photochemical ages that happen to be sampled in that particular study. The MS of the OOA-1 and OOA-2 factors are likely interpolants of the key variations in sources and/or age that cause spectral variation in a particular study. Therefore the OOA-1’s and OOA-2’s reported in different studies should not
be expected to be identical, reflecting differences in meteorology, transport time, and the mix of sources and precursors. Even for studies at the same location during different periods, these changes could lead to variations in the factor spectra.

The residual in the real dataset has considerable structure that changes very little with the addition of more factors, in stark contrast to the residual of the synthetic datasets, where the residual reflects only the noise added to the dataset when enough factors are chosen (Figs. 2.7, D.8). Similar results are reported in other component analyses of AMS data (Lanz et al., 2007; Zhang et al., 2005a), so none of these studies is fitting all of the real structure in the dataset. This distinct structure in Q/Qexp in the real dataset may imply that, though three factors have explained as much of the data as is possible with a bilinear model, something is changing during these periods of high residual and rotatability of the factors (Fig. 2.10). We hypothesize that aerosol partitioning or processing results in continuously, non-linearly varying spectra that cannot be fit well with the bilinear model. Evaporation and condensation of the semi-volatile OOA-2 component of the organic aerosol may lead to slight changes in composition that cannot be fit well with a constant mass spectrum. These changes in spectra, especially when fresh SOA/OOA is important in a dataset, may be a key limitation of the retrievability of the components in Q-AMS spectra. This should be a topic of further research.

2.5 Chapter Conclusions

An organic mass spectral dataset from Pittsburgh in 2002 was analyzed by PMF. The behavior of PMF solutions with AMS data was characterized using several 2- and 3-factor synthetic datasets with realistic noise. Three factors are identified from the Pittsburgh data. Two factors (HOA and OOA-1) are similar to the factors identified by Zhang et al. (2005a). A third factor was identified as OOA-2, a semi-volatile, less-oxidized OOA whose time series correlates
strongly with those of ammonium nitrate and ammonium chloride. The solutions are repeatable at different random starts and bootstrapping analysis supports the robustness of the solution. Researchers are urged to make these analyses, as well as to make other appropriate comparisons, such as between the results from robust and non-robust modeling. There is no clear support for justifying solutions with more than three factors. Note that this does not mean that there are only three sources, but rather that sources with very similar spectra (e.g., gasoline and diesel engine emissions) cannot be separated in this analysis with UMR data. Any sources that can be approximated by linear combinations of the PMF factors are likely partitioning among the retrieved factors and are not retrievable separately. Solutions with more than 3 factors appear to “split” the existing factors, a characteristic observed in synthetic datasets when more factors were calculated than existed in the input. Additional factors make minor changes in the residual of the Pittsburgh case, but appear to primarily refit the same variation in the data. We hypothesize that the structure in the residual reflects continuous, non-linear changes in the OOA-2 spectra as the aerosols partition or age, which cannot be fit by the bilinear model.

While the determination of a “best” solution is subjective and challenging in a real dataset, measures can be taken to make this process more quantitative. Correct specification of estimated error values, $\sigma$, for the dataset help prevent nonsensical factors (e.g., MS with one dominant fragment, TS that oscillate between zero and several $\mu$g/m$^3$ over short time intervals). Real and synthetic data indicate that plots of $Q/Q_{\text{exp}}$ vs. number of factors give a good indication of the minimum number of factors, but are not a sufficient criterion for choosing the best number of factors. Multiple random seeds should be tested to explore the possibility of local minima in the $Q$ space, and bootstrapping should be used to evaluate the statistical uncertainty of the candidate solutions. Max(RotMat) does not give any useful indication of the best number of
factors. Interpretation of the PMF solutions should start with the factor profiles (here, mass spectra). We have created a public, web-accessible database of AMS spectra which can be compared to spectra from PMF results to help identify and name the factors. Spectral similarity is not sufficient for naming factors, as “mixed” or “split” factors can have high similarity with many spectra in the database. Thus, correlations of time series with species not included in the PMF analysis are critical to give additional evidentiary support to a selected solution and interpretation. Best identification is made from supporting evidence for both mass spectra and correlation with tracer time series. It is imperative that both mass spectra and time series for each factor be presented because they are interdependent. Use of diagnostic plots showing the correlation between the factors in each solution is encouraged. It is unlikely that components with 5% or less of the mass are meaningful for Q-AMS data. Solutions with two factors may over- or underestimate the fraction of OOA, and this behavior appears to be dependent on the structure of the dataset. There is generally not a mathematical basis for choosing a particular rotation (using the parameter FPEAK), though correlation with external tracers and reference mass spectra may be used to narrow the set of plausible rotations. Presentation of solutions for several representative FPEAKs will allow readers to understand the rotational behavior and variability of the factors. A $\Delta Q/Q_{\text{exp}}$ over the minimum $Q/Q_{\text{exp}}$ of ~1% seems to be give appropriate range of FPEAKs for Q-AMS datasets.
Chapter 3

Three-dimensional factorization of size-resolved organic aerosol mass spectra from Mexico City

3.1 Chapter Introduction

Fine particles have important effects on human health, climate forcing, visibility, as well as deposition of nutrients and acids to crops and ecosystems. Some of the physical processes underlying these effects include deposition of toxic compounds into the lungs (Mauderly and Chow, 2008), cloud condensation-nucleus (CCN) activation (Andreae and Rosenfeld, 2008), scattering and absorbing of radiation (Watson, 2002), and particle settling and deposition (Feng, 2008). The extent and impact of these effects depends on both particle size and chemistry, with particles with submicron diameters being especially important. Aerosol interactions with radiation and clouds represent the greatest uncertainty of climate forcing (IPCC, 2007), and reducing this uncertainty requires a much better understanding of aerosol sizes and chemistry.

These aerosol effects are complex because aerosol size distributions are dynamic. Many processes can change the sizes of aerosols, including creation of new particles by nucleation; growth by coagulation and condensation; decrease in size by evaporation of semivolatile species upon dilution, heating, or chemical reaction; and removal of particles by wet or dry deposition (Whitby, 1978). Thus, measured ambient size distributions do not reflect the original sources directly, but may have been significantly transformed by atmospheric processes. Many of the
processes that change a particle’s size result in simultaneous changes to its chemical composition. For example, when semivolatile organic compounds condense onto an ammonium sulfate particle, the particle size increases, and the particle becomes an internal mixture. Thus aerosol size and chemical composition are directly linked, and ideally should be measured simultaneously.

Determination of size-resolved aerosol composition is challenging and is possible with only a few techniques. The traditional method for measuring size-resolved chemistry is to collect particles separately for discrete size ranges and chemically analyze the size-fractioned samples offline. Particles are segregated by momentum (which is proportional to aerodynamic size) and impact onto sampling stages (e.g., DRUM sampler, Cahill et al., 1987; MOUDI sampler, Marple et al., 1991). These methods usually have a time resolution of a few hours to days, limited by the minimum amount of material that needs to be collected for analysis, and labor and analysis costs, and cannot capture many dynamic changes of aerosol size distributions and chemical composition occurring over faster timescales.

In contrast, real-time measurements of size-resolved chemical composition can be made with aerosol mass spectrometry. Aerosol mass spectrometers can be divided into two main groups: those that measure single particles and those that measure the bulk aerosol. Many single-particle aerosol mass spectrometers measure the size and chemical composition of individual aerosol particles (PALMS, Murphy and Thomson, 1995; ATOFMS, Gard et al., 1997; Thomson et al., 2000; Su et al., 2004; SPLAT, Zelenyuk and Imre, 2005). The single-particle mass spectra collected by these instruments can then be clustered into types with similar composition and the size distribution of the cluster can be determined. These data give
qualitative results about the complete aerosol composition. In contrast, the Quadrupole Aerosol Mass Spectrometer (Q-AMS, Jayne et al., 2000) and Time-of-Flight Aerosol Mass Spectrometer (ToF-AMS, Drewnick et al., 2005; DeCarlo et al., 2006) quantitatively measure the size-resolved chemical composition of bulk submicron aerosol (~10,000 particles in several minutes). The Q-AMS monitors only a few ion fragments when sampling size-resolved chemical composition. In this mode, the quadrupole is stepped slowly across the selected ion fragments; consequently, only ions at one mass-to-charge ratio ($m/z$) are measured for each particle. In contrast, the ToF-AMS measures all ion fragments for each particle in this sampling mode to give a more complete and precise representation of the size-resolved chemical composition of the aerosol.

The chemistry of ambient inorganic aerosol is better understood because of its relative simplicity: only a few species are involved, and the set of possible reactions is limited. On the other hand, we have only begun to understand the complexity of organic composition, reactions, and volatility in the atmosphere. Organic compounds contribute ~20–70% of the submicron aerosol mass (Jimenez et al., 2009), so understanding their role in the atmosphere is important for understanding climate impacts. Organic aerosol (OA) enters the atmosphere by two mechanisms: it can be emitted directly (primary OA), or produced by secondary processes, including gas-phase chemical reactions produce low-volatility products that condense onto the surfaces of existing particles, and aqueous-phase reactions that produce species that stay in the particle phase (secondary organic aerosol, SOA). Most primary submicron organic aerosol is the result of incomplete combustion from sources such as vehicle engines and biomass burning. These sources, and vegetation, also emit SOA precursors. AMS measurements of submicron OA can provide insights into the contributions of different aerosol sources and processes. Aerosol components representing primary organic aerosol include hydrocarbon-like (HOA) and biomass-
burning organic aerosol (BBOA). Secondary organic aerosol is observed as oxygenated organic aerosol (OOA) in the AMS (Dzepina et al., 2009; Jimenez et al., 2009). These aerosol components have been identified by analyzing AMS measurements with a family of mathematical techniques known as factor analysis.

The purpose of factor analysis is to represent the measured data with a reduced number of physically meaningful factors that describe underlying sources and processes controlling the composition and variability of the original data. The possible factorization models depend on the contents of the data to be factored, how the data is arranged, and how the suspected causes underlying the data structure can be mathematically described. Factor analytical methods have been applied to chemically speciated aerosol data for more than forty years to identify the sources of particles in urban and rural environments (e.g., Blifford and Meeker, 1967; Thurston and Spengler, 1985; Chow and Watson, 2002; Engel-Cox and Weber, 2007; Reff et al., 2007; Viana et al., 2008). Size-resolved concentrations have also been included in some analyses, but only a few researchers have simultaneously factored chemical composition data from particles in different size ranges (Yakovleva et al., 1999; Dillner et al., 2005; Han et al., 2006; Pere-Trepat et al., 2007; Yatkin and Bayram, 2008; Amato et al., 2009; Gietl and Klemm, 2009; Karanasiou et al., 2009; Kleeman et al., 2009; Srivastava et al., 2009). All of these studies use data with limited size or temporal resolution. In fact, Larson et al. (2006) noted that their approach of combining number size distributions with bulk aerosol composition measurements in a factorization analysis was an interim approach until more size-resolved aerosol composition data from aerosol mass spectrometry was available. Size- and composition-resolved datasets such as those now available from the AMS have the potential to better characterize and constrain the changing size distribution of different aerosol species.
This study uses two 3Dimensional (3D) factorization models to analyze a size-resolved composition dataset collected during the intensive MILAGRO campaign (Megacity Initiative: Local and Global Research Observations) in Mexico City in March, 2006 (Molina et al., 2010). During the campaign, size-resolved submicron aerosol composition was measured near downtown Mexico City by high-resolution time-of-flight aerosol mass spectrometry (HR-ToF-AMS, hereafter AMS for short). Previous factor analysis of the bulk submicron composition dataset, i.e., without making use of the size-resolved data, identified contributions to the organic aerosol from HOA, BBOA, OOA, and a locally occurring organic aerosol (LOA, Aiken et al., 2010). The 3D factorization techniques used in this study allow us to obtain robust estimates of the size distributions of these aerosol components and perhaps identify additional components.

We first present a brief literature review of research that reports application of 3D factorization techniques to datasets that include aerosol size and composition information to determine sources of aerosol. We then discuss our methods, which include preparing the size-resolved aerosol composition measurements for factorization, and factoring these data using two 3D models with two algorithms. We present the results of the two factorization models, then compare the solutions and discuss the advantages and disadvantages of each for this dataset. Finally, we discuss the implications of these results for future studies of aerosol effects, and the use of these models for factoring 3D datasets from other emerging fast aerosol instrumentation.

3.2 3-dimensional matrix factorization and its application in the literature

3.2.1 Mathematical techniques for 3D matrix factorization

Factor analyses of 2-dimensional (2D) matrices produce factors composed of two vectors. Several approaches to factor analysis of datasets that include aerosol size distributions or size-
resolved aerosol composition datasets organized as 2D matrices are presented in Appendix H.1. In contrast, the third dimension in 3D matrices allows for more factorization model options (Fig. 3.1). Results from two models for the factorization of a few 3D size-resolved aerosol composition datasets have been reported in the literature and are summarized in Table 3.1. The first of these models is known as “Parallel Factor Analysis” or “PARAFAC” (Harshman and Lundy, 1994) and extends the bilinear umixing model [Eq. (H.1)] to a trilinear model (Fig. 3.1a). In this model, each factor is described by three 1-D vectors, and we refer to it here as the “3-vector model.” We arrange the \( m \times n \times o \) data matrix \( X \) of size-resolved chemical composition such that each of the \( m \) rows holds a 2D matrix of dimensions \( n \times o \) (columns \( \times \) layers of the 3D matrix) that contains the measurements of \( o \) chemical components measured at each of the \( n \) sizes for one sample (time step). Equivalently, the \( n \times o \) matrix could be thought of as the chemically resolved size distribution (across \( n \) size bins) of each of the \( o \) chemical components. The 3D data matrix \( X \) is reconstructed by a number of 3-vector factors as described by

\[
x_{ijk} = \sum_{p} a_{ip} b_{jp} c_{kp} + e_{ijk}, \tag{3.1}
\]

where \( i, j, \) and \( k \) are the row, column, and layer indices of a 3D matrix; \( p \) is the number of factors; \( a_{ip} \) is an element of the \( m \times p \) matrix \( A \), the columns of which contain the factor time series; \( b_{jp} \) is an element of the \( n \times p \) matrix \( B \), the columns of which contain the factor size distributions; \( c_{kp} \) is an element of the \( o \times p \) matrix \( C \), the rows of which contain the factor chemical profiles; and \( e_{ijk} \) is an element of the \( m \times n \times o \) matrix \( E \) of the residuals of the solution, i.e., the difference between the measured data and the reconstruction. Note that the 3-vector model incorporates the assumption that each factor’s size distribution and chemical composition is unchanging over the entire measurement period.
Fig. 3.1. Schematic representation of factorization methods of 3D matrices. (a) In the 3-vector model, each factor is represented by three vectors (here a mass spectrum, a size distribution, and a time series of the contribution of the factor). (b-d) In the vector-matrix model, each factor is represented by one vector (representing one of the original matrix dimensions) and one matrix (representing the remaining matrix dimensions). The model can be arranged so that any of the original dimensions is represented by the vector: (b) the vector represents a mass spectrum, (c) the vector represents a size distribution; (d) the vector represents a time series.
Table 3.1. Summary of research that applied 3D factorization techniques to datasets of size-resolved aerosol chemical composition.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Sampling Location</th>
<th>Instrumentation</th>
<th>Particle Sizes</th>
<th>Chemical Speciation</th>
<th>Sampling Time</th>
<th>Factorization Model(s)</th>
<th>Algorithm(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yakovleva et al., 1999&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Riverside, CA 1991</td>
<td>Stationary indoor monitors</td>
<td>PM&lt;sub&gt;2.5&lt;/sub&gt;, PM&lt;sub&gt;10&lt;/sub&gt;, personal PM&lt;sub&gt;10&lt;/sub&gt; (2 sizes, 5 types)</td>
<td>18 elements</td>
<td>12 hr</td>
<td>3-vector</td>
<td>PMF3&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Pere-Trepat et al., 2007</td>
<td>Detroit, MI 2002</td>
<td>3-stage DRUM impactor absorbance spectrometer</td>
<td>0.1–2.5 μm (3 ranges)</td>
<td>27 elements σ&lt;sub&gt;abs&lt;/sub&gt; (four wavelengths)</td>
<td>3 hr</td>
<td>Vector-matrix (time series as vector)</td>
<td>ME-2&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Karanasiou et al., 2009&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Athens, Greece 2002</td>
<td>Custom impactors aethalometer</td>
<td>PM&lt;sub&gt;2&lt;/sub&gt;, PM&lt;sub&gt;10–2&lt;/sub&gt;</td>
<td>SO&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;2-&lt;/sup&gt;, BC</td>
<td>24 hr</td>
<td>3-vector</td>
<td>PMF3</td>
</tr>
<tr>
<td>This work</td>
<td>Mexico City, Mexico 2006</td>
<td>HR-ToF-AMS</td>
<td>0.01–1.2 μm&lt;sup&gt;b&lt;/sup&gt; (36 ranges)</td>
<td>Organic fragments at &lt;i&gt;m/z&lt;/i&gt;’s ≤ 100 amu</td>
<td>5 min, 3 times per hr</td>
<td>- 3-vector - vector-matrix (mass spectrum as vector)</td>
<td>PMF3, ME-2</td>
</tr>
</tbody>
</table>

<sup>a</sup> In these studies, the data was also arranged as multiple 2-D matrices and factored using a 2-D model. These works are included in Table 3.1 and details are presented in Table H.5.

<sup>b</sup> Particle transmission in the AMS begins at ~0.04 μm, but smaller sizes were included in the factorization dataset to characterize noise level and provide a baseline for size distributions.

<sup>c</sup> Algorithms are Positive Matrix Factorization 3 (PMF3) and Multilinear Engine 2 (ME-2).
Another option for factoring 3D matrices produces factors composed of a 1-D vector and a 2D matrix. This model is often called the “Tucker1” model (Tucker, 1966), but we refer to it here as the “vector-matrix model.” Three such vector-matrix models are possible depending on which matrix dimension is chosen for the vector (Fig. 3.1b–d). The model in which the vector contains the chemical composition of one factor, and the matrix its time-varying size distribution (Fig. 3.1b), is described mathematically by

\[ x_{ijk} = \sum_p d_{ijp} c_{kp} + e_{ijk}, \]  

(3.2)

where \( x_{ijk} \) and \( e_{ijk} \) are as described as in Eq. (3.1); \( c_{kp} \) is the same as in Eq. (3.1) and is an element of the factor composition profile; and \( d_{ijp} \) is an element of the \( m \times n \times p \) matrix \( D \) in which each \( m \times n \) layer represents the time-evolving size distribution of each factor. In this model, the chemical composition of each factor is assumed to be constant at all times, but the size distribution of the aerosol component can change over time.

In contrast, the vector-matrix factorization could have two alternate arrangements. The vector-matrix model in which the vector contains the size distribution (Fig. 3.1c) could identify the changing composition of aerosol size modes. In the final vector-matrix model, the vector contains the time series, and the matrix represents size-resolved chemical composition (Fig. 3.1d). Thus each factor represents the size-resolved chemical composition of the ensemble of particles that arrive simultaneously at the sampling location.
3.2.2 Research reporting 3-dimensional factorizations using particle-size information

Three peer-reviewed studies have reported 3D factorization of size-resolved aerosol composition datasets using the 3-vector model, or the vector-matrix model in which the vector contains the time series. The datasets include data from the submicron range to PM10. Each is briefly discussed here.

Yakovleva et al. (1999) factored a dataset of particulate matter composition for particles with diameters less than 10 µm (PM10) and less than 2.5 µm (PM2.5) obtained during the 1991 Particle Total Exposure Assessment Methodology (PTEAM) study in Riverside, California. The goal of this study was to examine the relationship between daily activities and personal PM exposure. Particles were collected from five size-location combinations: PM10 samples were measured with personal, indoor, and outdoor monitors; and PM2.5 samples were measured with indoor and outdoor monitors. The samples were analyzed for concentrations of 18 elements. The dataset was factored using the 2D model (applied separately to each particle size range) and the 3-vector model (Fig. 3.1a). The 2D and 3-vector analyses identified factors with different contributions to the size-location combinations. For example, sea salt contributed primarily to outdoor PM10 with small contributions to outdoor PM2.5, while factors representing motor vehicle emissions and secondary sulfate were identified in all samples. In addition, particles generated by personal activities such as cooking, smoking, and vacuuming were identified in personal and indoor PM10 samples. The main difference between the factors from the 2D and 3-vector models was the separation of soil, identified as one factor in the 2D datasets, into three factors in the 3-vector results. These three soil factors had identical chemical profiles and were
identified as ambient soil in outdoor PM$_{10}$ samples, resuspended soil in personal PM$_{10}$ samples, and indoor soil in personal and indoor PM$_{10}$ samples.

Karanasiou et al. (2009) collected a dataset of coarse (PM$_{10-2}$) and fine (PM$_2$) particle samples at three sites in Athens, Greece, in 2002. The chemical speciation included 13 elements, black carbon (BC), and SO$_4^{2-}$. The dataset was factored with the 2D model (applied separately to each particle size range) and the 3-vector model. Using the 3-vector model enabled identification of some factors that could not be identified from the 2D factorizations. The factors identified only in the 3-vector model include motor vehicle exhaust contributions in the coarse fraction and a second road dust factor in both fractions. For some factors, different chemical profiles were obtained for the coarse and fine fractions from the 2D and 3D factorizations. This result differs from the results of Han et al. (2006), who found that aerosol sources had very similar chemical profiles across different size ranges (Appendix H1.3).

Pere-Trepat et al. (2007) report the only study in the literature that used a vector-matrix model to factor a 3D size-resolved dataset of particle chemical composition. The dataset for this study was collected in Detroit, Michigan, on three stages of a DRUM sampler and included particles with diameters from 0.1–2.5 µm. The collected particles were analyzed for the concentrations of 27 elements. In addition, absorbances at four ultraviolet and visible wavelengths were included in the factorization matrix. The authors report that they attempted to use the 3-vector model for this dataset; however, the factor compositions were size dependent, and thus the assumptions of the trilinear model did not hold. The authors therefore factored their dataset using the vector-matrix model in which the vector contains the time series (Fig. 3.1d). This vector-matrix model finds the size-resolved chemical composition of a component with one
time series, i.e., each factor represents the ensemble of particles that arrive simultaneously at the monitor, but the chemical composition of those particles may vary with size, though not with time. For example, the factor attributed to industrial metal works has the largest contribution from Fe at all sizes, but Ca appears only in the large particles and S occurs only in the larger two of the three size bins.

To the best of our knowledge, no published studies have applied the vector-matrix model in which the vector contains chemical compositions to aerosol datasets. Furthermore, we know of no studies that have applied 3D factorization techniques to highly time- and size-resolved chemical composition data.

### 3.3 Methods

In this section we discuss the collection of size-resolved aerosol composition data using the HR-ToF-AMS, preparation of that data for factorization, and the factorization models and algorithms used in this study. Finally, we present the guidelines used to choose factorization solutions.

#### 3.3.1 Mexico City measurements during the MILAGRO field campaign

Data were collected during the MILAGRO field campaign in Mexico City, Mexico, in March 2006 (Molina et al., 2010). The campaign urban supersite (“T0”) was located at the Instituto Mexicano del Petroleo (IMP), 9 km NNE of the city center, near a combination of residential, commercial, and light industrial areas. A HR-ToF-AMS (referred to hereafter as AMS, Aerodyne Research Inc., Billerica, MA; DeCarlo et al., 2006) was located at T0, and data were collected from 10 to 31 March, 2006.
The sampling procedures for the AMS during the MILAGRO campaign have been described in detail by Aiken et al. (2009) and are briefly summarized here. We use the size distributions recorded in the V, i.e., single-reflectron-pass, mode of the AMS for this study. Because of multiplexing the ambient data with other experiments, such as heating with a thermal denuder (Huffman et al., 2009), size distributions were collected for 2.5 minutes each in two adjacent time steps, or “runs”, of every eight time steps. Thus, 5 minutes of every 20 minutes were spent sampling in this mode, giving 6 samples per hour.

### 3.3.2 AMS sampling of particle time-of-flight data

Each time step includes data collected by quickly alternating between two sampling modes: MS mode, in which the ensemble-average mass spectrum of all particles is obtained; or Particle Time-of-Flight (PToF) mode, in which the size-resolved mass spectra, i.e., chemically resolved size distributions, of particles are obtained (Jayne et al., 2000; Jimenez et al., 2003). In both modes, particles and gases in ambient air enter the instrument and are focused into a beam by an aerodynamic lens. The lens transmits with nearly 100% efficiency particles with a vacuum aerodynamic diameter ($d_{va}$) of 70–600 nm. It transmits with decreasing efficiency particles with $d_{va}$ 35–70 nm and 600 nm to 1.5 μm (Liu et al., 1995a, b; Jayne et al., 2000; Zhang et al., 2004). Gases are also transmitted through the lens with 100% efficiency, but are not focused at the exit of the lens. Particles and gases that are transmitted through the lens encounter a rotating chopper wheel (~100 Hz) that is positioned differently for the MS and PToF sampling modes. Only the PToF mode is described here. In PToF mode, the chopper is positioned to partially block the particle beam, allowing a few particles through the slit during each chopper cycle. Particles and gases that pass the chopper impact a resistively heated surface (600°C), and nonrefractory
aerosol components flash vaporize. The particle vapors and gas molecules are ionized (70 eV electron impact), and the resulting cations are transmitted by ion optics to an extraction region. In the extraction region, the ions are extracted into the ToFMS and fly to the detector.

Particle size is determined from particle velocity through the particle-flight region. Particle velocity is calculated from the length of the particle-flight region (0.293 m) and particle time of flight (PToF). PToF is measured as the time between the midpoint of the opening of the chopper slit intersecting the particle beam and the arrival of chemical signal at the MS detector. This PToF includes the true PToF from the chopper to the vaporizer, and ion time of flight (IToF) from the extractor to the detector; however, since IToF ≪ PToF (<30 µs and ~3000 µs, respectively), the approximation of PToF as measured is sufficient. The measured PToF and the length of the particle flight region give the particle velocity, which is inversely proportional to the square root of particle vacuum aerodynamic diameter ($d_{va}$) (Jayne et al., 2000; DeCarlo et al., 2004).

Size-resolved mass spectra are obtained by recording many individual mass spectra as a function of particle flight time (Fig. 3.2a). The frequency of recording the mass spectra determines the achievable size resolution. Mass spectra are recorded, e.g., 50 times during each chopper cycle. Most of the mass spectra contain only noise because zero to a few particles enter the particle flight region during each chopper cycle, so only a few spectra contain particle signal. Size-resolved mass spectra from ~10,000 chopper cycles during one 2.5-minute time step are averaged into one size-resolved mass-spectral matrix for that time step (Fig. 3.2b). Thus for a typical urban concentration of 10,000 particles/cm³ sampled in PToF mode during 50% of a 2.5-
**Fig. 3.2.** Schematic representation of data collection for chemically resolved particle-size-distribution data.  
(a) An example case is shown in which two particles have entered the particle flight region, before being vaporized and ionized. Mass spectra are collected ~50 times during the particle flight time (circles). The ions from the smaller particle arrive first (blue circle), and their mass spectrum is recorded (blue spectrum). The ions from the larger particle arrive second (green circle), and their mass spectrum is recorded (green spectrum). [Adapted from Cross et al. (2009) with permission.]  
(b) The size-resolved mass spectra can be arranged as a 2-D matrix in which one dimension represents the mass-to-charge (m/z) ratio, and one dimension represents particle size. The same data could be conceptualized as m/z-resolved size distributions if the points in each m/z row are connected.  
(c) Over many measurement periods, the 2-D samples can be arranged into a 3-D matrix.
minute time step with a flowrate of 2.0 cm$^3$/s and a 2% chopper opening, this matrix contains the average size-resolved mass spectra of ~30,000 particles.

### 3.3.3 Particle time-of-flight data analysis

Quantification of the size-resolved particle signal requires subtracting the background signal. However, the background signal is not explicitly measured in PToF mode. Instead, PToF-mode samples (time steps or “runs”) include measurements before and after particle signals arrive at the detector, and these periods, or “DC regions,” can be used to estimate the signal background (Allan et al., 2003). The signal from both periods is usually used to determine the background (Fig. H.1a); however, some ions have high signal from gas-phase ions that arrive before the particles, and only the latter region is used to determine the background level (Fig. H.1b). The signal over the selected regions is averaged, and, analogous to a "DC offset", is subtracted from the measured signal. The background subtraction is performed separately for each ion in every time step.

After subtracting the background from the PToF matrix of the total aerosol and gas-phase signal, the organic ion fragments are separated from the bulk aerosol by applying a "fragmentation matrix" to the mass spectrum measured at each particle size (Allan et al., 2004). Details of changes to the standard fragmentation matrix for this study are included in Appendix H.2.

We now arrange the size-resolved organic aerosol mass spectra obtained over many sampling periods into a 3D matrix of dimensions $m \times n \times o$, as described in Sect. 2.3 (Fig. 3.2c). The matrix is arranged so that the rows represent time steps ($m$ time averages), the columns
represent particle size (n sizes), and the "layers" or "pages" (the third dimension) represent ion fragments in a mass spectrum (o m/z’s, Fig. 3.1). Each element of the matrix has units dSignal/dPToF, which are nonlinearly transformed to units of dSignal/dlogdva (Allan et al., 2003). Units of dSignal/dlogdva are converted to dMass/dlogdva by integrating dSignal/dlogdva for each time step, applying AMS calibration factors (Jimenez et al., 2003), and normalizing the total PToF signal to the total mass of the same time step measured in MS mode.

3.3.3.1 Estimation of measurement precision of particle time-of-flight data

A method to estimate the precision of AMS PToF data has not been reported previously. This quantity is required for the factorization analyses, and so we have developed a method for its estimation here. The precision (or random error, σ, often termed “error” in studies using Positive Matrix Factorization, PMF) of the measured signal for one m/z at one size in each time step can be estimated by the sum of three terms in quadrature:

\[
\sigma = \sqrt{\sigma_{IC}^2 + \sigma_{DC}^2 + \sigma_{elec,scat}^2},
\]

(3.3)

where \(\sigma_{IC}\) is the precision from Poisson ion-counting statistics, calculated from the ion signal before DC offset subtraction [i.e., \(\sigma_{IC} = \sqrt{I/t_s}\), where \(I\) is the ion signal in ions per second and \(t_s\) is the time spent sampling that size bin and m/z in seconds, analogous to the method of Allan et al. (2003) for MS mode data]; \(\sigma_{DC}\) is the standard error of the signals used to estimate the DC offset (estimated as \(\sigma_{DC} = s/\sqrt{n}\), where \(s\) is the sample standard deviation of the points averaged for the DC offset and \(n\) is the number of points in that average); and \(\sigma_{elec,scat}\) is the error from electronic noise and scattered ion signals that are presumed to contribute to every measured ion signal, estimated as the standard deviation of the signal at very high m/z’s at all particle sizes. In
this study, $\sigma_{\text{elec,scat}}$ was estimated from the signals at $m/z$’s > 400 for 200 time steps that did not appear to contain actual particle signal (defined as signal at least four times the average noise level).

3.3.3.2 Further data and error treatments prior to factorization

Further treatments to the data and error matrices are needed prior to matrix factorization for three main purposes: (1) to decrease the influence of low signal-to-noise ratio (SNR) data in the factorization, (2) to remove matrix elements that do not contain useful particle information, and (3) to remove the weight of duplicated information within the matrix. To implement these treatments, we follow the five-step procedure described by Ulbrich et al. (2009), with some additions specific to the PToF data.

The first step in the Ulbrich et al. (2009) procedure is to calculate the error for each measurement according to Eq. (3.3). In the second step, we apply a minimum error value equal to the signal measured from one ion during one time step. This correction is applied because the uncertainty of counting cannot be less than one count, but the values calculated from Eq. (3.3) may be smaller than one count. In the Ulbrich et al. (2009) study, this step had two functions: it replaced error values calculated as less than one ion, but also identified error values that were lower than the average of its neighbors in time and replaced the calculated error with that average. However, this second part of the correction cannot be made here because only two adjacent measurements were made, so no point has two truly contiguous neighbors. The minimum-error correction increases the estimated $\sigma$ only for very small signals. In this study, one ion is equivalent to 1.24 Hz or 0.03 $\mu$g/m$^3$/decade log($d_{\text{ua}}$), and 52% of the data points included in the final matrix have their error increased in this step by ~8% on average. The
increases are larger for particular m/z’s and for some particle sizes (Fig. H.2). Many ions with m/z > 85 required uncertainty increases of 20–70%. The average increased uncertainty at each particle size was of smaller magnitude than the average increase by m/z and mainly affected particles that are transmitted with lower efficiency through the aerodynamic lens, i.e., \(d_{va} < 45\) nm or \(> ~900\) nm. The large fraction of points in the matrix that require this correction is a consequence of the limited SNR for this PToF dataset.

The limited SNR of the PToF-mode data is partially the result of lower sampling duty cycle in PToF mode than in MS mode. In MS mode, the particle beam is sampled at least 50% of the time during the “open” mode, but in PToF mode, the chopper allows particles to enter the sampling region during typically 2% of the sampling time. To reduce the impact of high-frequency noise on the PToF data, we follow the third step of the Ulbrich et al. (2009) procedure and smooth the data in the size and time dimensions. The particle size dimension is smoothed binomially by two points for each m/z in each time step. In the time dimension, two adjacent time steps were averaged together to obtain 5-min sample averages. The effect of these steps in the estimated precision was propagated in the PToF uncertainties.

The minimum error step shows that many of the signals in the matrix are small. In fact, the matrix includes many points that we know do not contain useful information about particle size or composition. In the size dimension, some points represent signal before and after particles can arrive; indeed, these are the sizes that were used to calculate the PToF background in Sect. 3.3. Because these data should not contain particle signal, we discard them and retain the data for nominal particle sizes \(10 \text{ nm} \leq d_{va} \leq 1200\) nm. Even at the extremes of this range, we expect very little or no particle signal, but we retain these edges so that both tails of the size
distributions approach zero. We have thus removed data points that do not contain aerosol size information and can now examine \( m/z \)'s with little chemical information. The signal from organic fragments is found predominantly at \( m/z \)’s \( \leq 100 \). For example, DeCarlo et al. (2008) report that 91% of the organic signal was found below \( m/z \) 100 for an aircraft HR-ToF-AMS dataset during MILGRO. In addition, ions at higher \( m/z \)’s have lower SNR. Mass fragments with \( m/z \)’s from 1 to 826 were measured in PToF mode during this study, but for this analysis we retain only \( m/z \)’s \( \leq 100 \) with organic fragments. Finally, 2% of the time steps have total PToF mass concentration less than zero because of noise at low actual concentrations, and we omit these time steps from further analysis. These steps for removing particle sizes, ion fragments, and time steps with little organic particle information were not part of the Ulbrich et al. (2009) procedure.

After removing sections of the matrix with little organic particle information, the resultant matrix contains time steps at 1366 times (rows), 36 size measurements (columns), and 71 \( m/z \)'s (layers), or \( \sim 3.5 \times 10^6 \) data points. This is a 32-fold reduction from the original matrix size of \( \sim 1 \times 10^8 \) data points. Using this smaller matrix for the factorization analysis greatly reduces computer time, memory, and storage requirements while preserving the high time-, size-, and chemical resolution of the useful information.

**Identifying and downweighting data with low signal-to-noise ratio**

Now that the data matrix contains only the points that will be used for further analysis, we assess the SNR to identify portions of the data with low-information content. The data with low-information content are reweighted, or “downweighted”, to decrease their weight in the fit (Paatero and Hopke, 2003).
Metrics for assessing the information content of the data based on calculated SNR are discussed in detail by Paatero and Hopke (2003) for 2D datasets, but these authors make no recommendations for 3D datasets. In 2D datasets, average SNR is calculated for each variable over all of the time steps; the variables in 2D AMS datasets are the \( m/z \)’s. In contrast, the average SNR for 3D datasets can be calculated in more ways. Three examples are given here: (1) the SNR could be calculated for each \( m/z \) at each size, averaged across all of the time steps; (2) the SNR could be calculated for each size, averaged across all of the \( m/z \)’s and time steps; or (3) the SNR could be calculated for each \( m/z \), averaged across all sizes and time steps. The third method was used in the only published 3D factorization study that downweighted low-SNR data (Pere-Trepat et al., 2007). In that study, the authors reweighted the data for selected chemical species at all sizes and all times by factors of 3 or 10, but no details were provided about the criteria used to determine which species should be downweighted. In our dataset, however, we know from applying the minimum error that the information content (SNR) in the matrix varies with \( m/z \) and particle size. For example, the minimum error was much larger than the calculated error at many of the higher \( m/z \)’s, and the smallest and largest particle sizes contain little particle information regardless of \( m/z \). Thus, we use the first method and calculate the SNR separately for each \( m/z \)-size combination.

We now use the average SNR values to downweight \( m/z \)-size combinations with low SNR; this is the fourth step of the Ulbrich et al. (2009) error preparation procedure. The threshold for low SNR recommended by Paatero and Hopke (2003) is 2, and the recommended threshold for very-low SNR is 0.2; Paatero and Hopke call the low and very-low SNR data “weak” and “bad,” respectively. No data are bad in the current dataset by that criterion. However, 89% of the \( m/z \)-size combinations are weak when using the recommended threshold of
SNR < 2. Since we have established that the SNR is generally weak for this data, and Paatero and Hopke note that the SNR threshold for “weak” variables is somewhat arbitrary, we set the threshold for weak data at SNR of 1.5. With this change, 76% of the m/z-size combinations are still weak, but nearly all m/z’s have strong signal for particles sizes that are transmitted through the instrument’s aerodynamic lens with 100% efficiency (Fig. H.3). We can therefore emphasize these high SNR data by downweighting the weak m/z-size combinations. We increase the calculated error for the weak m/z-size combinations by a factor of 2.

We must address one additional set of points that does not contain useful data about particles: data points that have interferences from gas-phase species. Recall that after gases and particles enter the AMS, both pass through an aerodynamic lens (Sect. 3.3.2). The lens focuses the particles, but not the gases; however, about 1 in 10⁷ gas molecules travels along the axis of the chamber and is ionized. Because the gases have a velocity distribution related to the Maxwell-Boltzman distribution and travel at higher velocities than the particles, we measure ions from these molecules both at times before we expect particle signal (from most of the gas-phase molecules), and also when we would expect signal from extremely small particles (from the slowest gas-phase molecules). The time variation of these gas molecules’ signal is very different from the particle signals, and this pattern will be identified in the factorization unless the interfering gas molecule signal is deemphasized by downweighting. This step was not part of the Ulbrich et al. (2009) procedure.

Potential interfering ions are due to the major components of air (N₂, O₂, H₂O, Ar, and CO₂). The four main constituents (N₂, O₂, H₂O, and Ar) have already been removed completely in the default fragmentation matrix because no organic signal is assigned to m/z’s 28, 32, and 40.
The remaining interferences arise from N\textsuperscript{15}N at m/z 29 and CO\textsubscript{2} at m/z 44. The signal from these gases is estimated from their abundance in ambient air relative to N\textsubscript{2}. Specifically, N\textsuperscript{15}N is estimated from the isotopic abundance of \textsuperscript{15}N, and the average concentration of CO\textsubscript{2} is also known. The PToF-mode signal of these gases is assumed to have the same shape, i.e., distribution in PToF, as that of N\textsubscript{2}, and the apparent size distributions for N\textsuperscript{15}N and CO\textsubscript{2} are subtracted from the measured size distribution at m/z’s 29 and 44 with the fragmentation matrix. However, the relationship of the apparent size distributions of N\textsubscript{2}, N\textsuperscript{15}N, and CO\textsubscript{2} has not been studied in great detail and the assumed subtractions have greater uncertainty than is estimated from the application of the fragmentation table. We therefore downweight the signal for both m/z 29 and m/z 44 at smallest nominal particle sizes quite strongly (by a factor of 100), then decrease the downweighting toward larger nominal particles sizes as the expected tail of the gas-phase interference decreases (Table H.1).

Finally, we perform the last step of the Ulbrich et al. (2009) procedure and downweight information that is repeated in the data matrix because of the application of the fragmentation matrix. In the fragmentation matrix, the organic signal and uncertainty at m/z’s 16, 17, and 18 are defined to be proportional to the signal at m/z 44. Thus the information for m/z 44 is repeated in the data matrix four times. This repetition is chemically meaningful, but if these m/z’s are used without modification, they have undue additional weight in the factorization analysis. We therefore downweight the signal at these four m/z’s by the square root of 4 so that this information is weighted the same as any other single m/z (Ulbrich et al., 2009).
3.3.4 Matrix Factorization

After performing these steps, the data matrix has been prepared for factorization. We now discuss the factorization models used in this study, the algorithms used to solve the models, and conclude by presenting the guidelines used for choosing factorization solutions. All factorization results were examined using the PMF Evaluation Tool (PET) described previously (Ulbrich et al., 2009), with custom modifications for 3D matrix factorization.

3.3.4.1 Models for factoring the 3-Dimensional matrix

Four models for factoring 3D matrices were presented in Sect. 3.2 (Fig. 3.1). This study applies two of these models to the MILAGRO dataset. The first model is the 3-vector model (Fig. 3.1a). Recall that the 3-vector model is so named because each factor is composed of three vectors. When this model is applied to the present dataset, the vectors contain the factor’s chemical composition (here a mass spectrum), size distribution, and mass concentration time series. The 3-vector model thus assumes that each factor’s size distribution and chemical composition is unchanging over the entire measurement period. The second model used for this study is the vector-matrix model. Recall that in the vector-matrix model, each factor is composed of a vector and a matrix. In this study, we use a vector-matrix model in which the vector contains the aerosol composition, i.e., mass spectrum (Fig. 3.1b). Hereinafter we use “the vector-matrix model” to refer to this variant of the three possible vector-matrix models, unless otherwise noted. This vector-matrix model assumes that each factor’s chemical composition is constant at all times, but the size distribution of each factor can change over time. To compare the factors from solutions of the 3-vector and matrix models, we must match the shapes and units. In the factors from both models, we normalize mass spectra to sum to 1 [as is standard for
AMS data (Ulbrich et al., 2009), but in contrast to the usual mass spectrometric practice of normalizing the $m/z$ with the highest signal to 100]. In the 3-vector model, we normalize the area under the size distribution to sum to 1, thus giving units of mass concentration ($\mu g/m^3$) to the time series. In contrast, in the vector-matrix model, the matrix represents $dM/d\log d_{va}$ with units $\mu g/m^3$/decade $\log(d_{va})$. From this matrix, a factor’s average normalized size distribution can be calculated by averaging all of the size distributions in the matrix, then normalizing to unit area. Similarly, the total time series can be calculated by summing the area under the size distribution from each time step. These two reductions of the matrix from the vector-matrix model can then be compared directly with the 3-vector model results.

Solving these models requires no a priori information about the factor mass spectra, size distributions, or time series. However, a priori knowledge can be incorporated to constrain the solution if the researcher deems the additional information appropriate and reliable. For selected analyses in this study we use the mass spectra obtained from the previous analysis of this dataset (Aiken et al., 2009; henceforth referred to as "the HR-MS" factors, spectra, or solution) as starting guesses for solving the models. Thus, a priori information is available, but it may not directly correspond to the mass spectra obtained from factoring the PToF data. For example, different evaporation timescales of different organic components on the AMS vaporizer could lead to slightly different factors obtained from the mass spectral data compared to the characteristic mass spectra recorded in the PToF mode.

To allow for differences between the HR-MS and PToF mass spectra, we introduce a parameter that allows the intensity, $c$, at each $m/z$ in each reference factor to deviate from its starting value. Lanz et al. (2008) used a parameter, $\alpha$, that allowed $c$ in an a priori mass spectrum
to vary by a fraction $\pm \alpha$ from its original value. With this parameter, the allowed range from the starting value $c_0$ is $c_0 - \alpha c_0$ to $c_0 + \alpha c_0$, and $\alpha$ may have values from 0 to 1. Thus when $\alpha$ is at its maximum, $c$ may range from 0 to $2c_0$. However, if an a priori spectrum is too dissimilar from the latent spectra in the dataset — especially if the starting guesses have too-small values for important ion fragments — even an $\alpha$ of 1 may not allow the solution sufficient flexibility to find a good solution. Thus we define a different parameter, $\beta$, which allows $c$ to fractionally approach the limits 0 and 1. We implement $\beta$ by

$$
\begin{align*}
  c_{\text{low}} &= c_0 - \beta(c_0 - 0), \\
  c_{\text{high}} &= c_0 + \beta(1 - c_0),
\end{align*}
$$

(3.4)

where $c_{\text{low}}$ and $c_{\text{high}}$ are the low and high limits for $c$, respectively. In this formulation, $c_{\text{low}}$ is identical to its formulation by $\alpha$, but $c_{\text{high}}$ allows $m/z$'s with small $c_0$ to grow substantially if necessary.

### 3.3.4.2 Algorithms for solving the 3-Dimensional models

The two algorithms/software tools used in this study to solve the 3-vector and vector-matrix models are Positive Matrix Factorization 3 (PMF3, Paatero, 1997) and the Multilinear Engine 2 (ME-2, Paatero, 1999). PMF3 can only solve the 3-vector model, while ME-2 is a flexible tool that can solve both models, as well as other multilinear and quasi-multilinear models (Paatero, 1999). Both algorithms constrain the values in the factor matrices to be positive. Specifically, in PMF3 the values in factors are constrained to be positive (Paatero, 1997), and in ME-2 the values of the factors are constrained by default to be non-negative, i.e., $\geq 0$ (Paatero, 1999). The positivity constraint helps produce physically meaningful factors.
because real mass spectra, size distributions, and mass concentrations have all positive values, and negative values arise only because of noise.

Both algorithms evaluate potential solutions by minimizing a quality of fit parameter, $Q$, defined as the sum of the error-weighted residuals of the entire data matrix, or

$$Q = \sum_{i=0}^{m} \sum_{j=0}^{n} \sum_{k=0}^{o} \left( \frac{e_{ijk}}{\sigma_{ijk}} \right)^2. \quad (3.5)$$

A theoretical “expected” value of $Q$, ($Q_{exp}$) is approximated by the number of points in the data matrix minus the degrees of freedom in the solution (i.e., the number of points in the solution matrices, Paatero et al., 2002), i.e.,

$$Q_{exp} = mno - p(m + n + o) \quad \text{for the 3-vector model, and} \quad (3.6)$$

$$Q_{exp} = mno - p(mn + o) \quad \text{for the vector-matrix model.} \quad (3.7)$$

The change in $Q_{exp}$ with the addition of each factor (i.e., increasing $p$ by 1) is small for the 3-vector model, but more substantial for the vector-matrix model. For example, in the present dataset, each additional factor in the 3-vector model decreases $Q_{exp}$ by 0.04% of the matrix size $[(m + n + o)/mno]$, while each additional factor in the vector-matrix model decreases $Q_{exp}$ by 1.4% of the matrix size $[(mn + o)/mno]$. Because the vector-matrix model has more degrees of freedom, we expect to fit substantially more of the matrix information with the vector-matrix model than the 3-vector model.

We use $Q_{exp}$ to normalize the $Q$ values for solutions of the models. If each point in the matrix is fit within its prescribed uncertainty, $e_{ijk}/\sigma_{ijk}$ is $\sim 1$, $Q \sim$ the size of the matrix ($mno$), and
$Q/Q_{exp} \sim 1$. However, if the uncertainty values have been calculated incorrectly, $Q/Q_{exp}$ may be higher or lower than 1. Note that $Q/Q_{exp}$ may be larger than 1, even if the errors have been specified correctly. Increased $Q/Q_{exp}$ values can result from variations in the data that do not behave according to the model. For example, if a factor’s size distribution is not constant, it will not be fit well with the 3-vector model. Similarly, if a factor’s mass spectrum varies in time in a way that cannot be fit well with an additional factor, it will be fit poorly by both models (Ulbrich et al., 2009). The $Q/Q_{exp}$ values calculated with the error matrix used in the computations in this study are artificially low since so many of the points have been downweighted because of low SNR, interference from gas-phase molecules, or repetition of chemical information. Thus we find it useful to recalculate the $Q/Q_{exp}$ values using the error estimates calculated before downweighting (these error estimates do include application of the minimum error and uncertainty propagation of smoothing). We call these recalculated values “unweighted $Q/Q_{exp}$ values.” We present unweighted $Q/Q_{exp}$ values in this study unless otherwise noted.

To solve the models, both the PMF3 and ME-2 algorithms begin by filling the factor matrices with random values determined from random seeds. Then the algorithms iteratively minimize $Q$. The two algorithms use different minimization algorithms. PMF3 uses a Gauss-Newton algorithm (Paatero, 1997) while ME-2 uses the conjugate gradient method (Paatero, 1999). The difference in minimization algorithms means that PMF3 and ME-2 may not find identical solutions for the same problem, though solutions from the two algorithms should be similar.

Finally, we note several details about the configuration of the algorithms for this study. We run both algorithms in the "robust mode," in which outliers ($|e_{ijk}/\sigma_{ijk}| > 4$) are dynamically
Reweighted during the iteration so that they cannot pull the fit with weight > 4. The algorithm uses three levels of convergence during the iteration; the convergence criteria for these levels were set to $Q_{exp} \times 10^{-4}$, $Q_{exp} \times 2\times 10^{-5}$, and $Q_{exp} \times 10^{-5}$, respectively. Thus the final convergence criteria were ~35 and ~31 absolute $Q$-units for the 3-vector and vector-matrix models, respectively.

### 3.3.4.3 Guidelines for choosing a solution

The solutions to positively constrained unmixing models are not unique, and no set of mathematical criteria have proven sufficient to identify the best solution of a factor analytical model. Thus the modeler must choose the “best” solution from the set of possible solutions; this choice is unavoidably subjective. Paatero and Hopke (2009) note the importance of disclosing subjective decisions in publications of factor analyses so that the analyses can be repeated or altered by other researchers. In that spirit, we present here our guidelines for choosing the best solution from the 3D factorizations, and discuss the acceptability of each candidate solution of the two models in two Appendices to the paper.

Choices regarding the best solution must be made in two main areas. First, the number of factors in the solution must be determined. Second, multiple solutions with this number of factors may exist, and one family must be chosen as the best solution. The choice of the best number of factors is discussed in more detail elsewhere (Paatero and Tapper, 1993; Paatero et al., 2002; Ulbrich et al., 2009) and is only briefly described here. We use the recommendations of Ulbrich et al. (2009) and consider these criteria for choosing a solution: $Q/Q_{exp} \sim 1$, decrease in the rate of change of $Q/Q_{exp}$ with increasing number of factors, little structure in the solution residuals, strong correlation between component time series and diurnal cycles with those of
tracers not included in the factorization matrix, and plausibility of the factor mass spectra and their similarity to observed spectra of real-world sources. For this dataset, we can also compare to the HR-MS solution, which identified four factors: HOA, BBOA, OOA, and LOA. We presume that the bulk aerosol has the same composition whether measured in the MS and PToF modes since the sampling modes are alternated every few seconds, i.e., much faster than aerosol sources or processes change. We therefore hypothesize that we should find the factors identified in the HR-MS analysis in the 3D analysis, and may identify additional factors.

It is possible that the solution space for a given number of factors may contain multiple solutions; these solutions represent local minima in the $Q$ space (Paatero, 2000, 2007). Solutions from different local minima usually have different $Q/Q_{exp}$ values, but the solution with the lowest $Q/Q_{exp}$ value is not necessarily the best solution. The possibility of solutions at local minima can be explored by varying the seed for the starting of the algorithm so that the algorithm begins from different parts of the $Q$ space and might therefore encounter local minima. We calculate each solution from 50 different starting seeds and then compare these solutions (Ulbrich et al., 2009). The 50 solutions can be grouped into “families” of solutions by comparing $Q/Q_{exp}$ values and the similarity of factors within the family (Allan et al., 2010; DeCarlo et al., 2010). Each family can then be represented by the average of the solutions in that family, and one family can be selected as the best solution.

### 3.4.4 Uncertainties in the chosen solution

In addition to choosing the best solution, we would like to quantify the uncertainty of the factors in this solution. Three main approaches to estimating the uncertainty of PMF and ME-2 solutions have been reported. First, PMF3 can report the standard deviations of the elements of
one factor matrix \([\mathbf{A}, \mathbf{B}, \mathbf{C}]\) in Eq. (3.1)] while the elements of the others are held fixed, e.g., standard deviation of \(\mathbf{A}\) when \(\mathbf{B}\) and \(\mathbf{C}\) are fixed (Paatero, 2007). However, the quantitative nature of these estimates has not been proven (Paatero, 2007), and the estimates appear to be too small in our experience. Furthermore, such estimates are not reported by ME-2 and could not be compared for the vector-matrix model; therefore we do not use this method. Second, bootstrapping with row replacement (Press et al., 2007; Norris et al., 2008; Ulbrich et al., 2009) has been used with 2D input matrices to estimate uncertainty in the factors; however, for 3D input matrices more replacement schemes are possible and exploration of this complex topic is outside the scope of the current study. Finally, uncertainty in the factors can be estimated by the variation in solutions in the same family. The variation amongst these solutions may depend strongly on the shape of the \(Q\) space near a local minimum and the algorithm’s convergence method. Therefore these variations may better describe the solutions’ mathematical variation than the physical variation that would help us evaluate uncertainty in aerosol properties. Nevertheless, this approach may give some insight into the mathematical uncertainties of the solution. We calculate the variation among the solutions as the coefficient of variation \((\sigma/\mu\), where \(\mu\) is the mean) of average mass spectra over all \(m/z\)’s, size distribution over all size bins, or time series over all times, after excluding points with very small means (below 0.002 fraction of mass-spectral signal, 0.006 \(\mu g/m^3/\text{decade}\ \log(d_{\text{ua}})\) for size distributions, and 0.005 \(\mu g/m^3\) for time series).

3.4 Results

This section presents the results of the 3-vector and vector-matrix model factorizations, including the choice of the “best” solution for each model. We first present results that apply to
both models, then discuss the choice of solution and physical interpretation of the factors for each model. Several results apply broadly to factorization with both the 3-vector and vector-matrix models for the current dataset and are discussed here.

\(\frac{Q}{Q_{\text{exp}}}\) values calculated by the algorithms using the downweighted error matrix were much smaller than the expected value for a good solution (~0.45 vs. 1, Fig. 3.3). The low \(\frac{Q}{Q_{\text{exp}}}\)

![Fig. 3.](image)

**Fig. 3.** Values of \(\frac{Q}{Q_{\text{exp}}}\) for the 3-vector and vector-matrix factorization models with 1 to 8 factors. Two algorithms were used to solve the two models: PMF3, which can only solve the 3-vector model; and ME-2, which can solve a variety of factorization models. Each mark shows the \(\frac{Q}{Q_{\text{exp}}}\) value of one of 50 seeds used to start the algorithm. Lines connect the average \(\frac{Q}{Q_{\text{exp}}}\) value of the 50 seeds for each model-algorithm combination. Marks that appear as an “×” are the intersection of “/” and “\". In the constrained vector-matrix case (green dashes), the mass spectra of four factors are fully constrained as described in the text. (a) \(\frac{Q}{Q_{\text{exp}}}\) values have been recalculated without downweighting. (b) \(\frac{Q}{Q_{\text{exp}}}\) values shown were calculated with the downweighted error matrix.
values reflect the large fraction of size-bin–$m/z$ combinations that were downweighted because of low SNR. However, unweighted $Q/Q_{exp}$ values are close to 1 for solutions of both the 3-vector and vector-matrix models (Fig. 3.3a). We would expect unweighted $Q/Q_{exp}$ values to be somewhat greater than 1 unless the model can fit all the real variability in the data. Or, the errors may be slightly overestimated for the PToF data. We expect that any small, systematic issues in our error estimate procedure do not depend strongly on $m/z$, size, and/or time. Thus, the weighting of the data in this study should be consistent and we believe that our results represent the actual structure of the dataset.

Another suggested criterion for choosing a solution for factor analytic models is a steep change in the slope of $Q/Q_{exp}$ vs. the number of factors in the solution. No such change was observed in any of the models for this dataset; thus we cannot use this criterion to choose a solution and must rely upon the remaining criteria from Sect. 3.3.3.3: little structure in the solution residuals, correlation with the time series of tracers, and identification of the factors from the HR-MS solution.

Taking these criteria into account, we explore solutions of the models. In solutions with at least three factors, the 50-seed solutions do not all contain the same factors, but the solutions can be arranged by factor similarity into several families (Table 3.2). Although seed-dependent families have been reported previously for 2D factorization of some AMS datasets (Allan et al., 2010; DeCarlo et al., 2010), our solutions within the same family show less variation than in the reported 2D cases. The variation among the solutions in each family is quite small. In the 3-vector solutions solved by either algorithm, the average coefficient of variation of the mass spectra, size distributions, or time series of a family is less than 2%, in the unconstrained vector-
Table 3.2. Types of factors identified in each solution family from 50 “seed” trials of the 3-vector and vector-matrix models solved by ME-2. Each row represents one family of solutions, and each X represents one factor in a solution. Multiple Xs in one entry denote multiple instances of this factor in the same solution. These factors usually have similar mass spectra but different size distributions and time series (Fig. H.6). The columns with darker grey shading represent factors related to oxygenated organic aerosol (OOA). Each solution contains one OOA factor or two factors dominated by \( m/z \) 44 and \( m/z \) 43. The columns with lighter grey shading represent factors that are dominated by \( m/z \) 15 or \( m/z \) ’s 67, 81, and 95. These factors are not physically meaningful. The best solution of the 3-vector model has four factors and is shown in bold. No solutions of the vector-matrix model were considered physically meaningful.
<table>
<thead>
<tr>
<th>Number of Factors in Solution</th>
<th>OOA</th>
<th>$m/z$ 44</th>
<th>$m/z$ 43</th>
<th>HOA</th>
<th>BBOA</th>
<th>LOA</th>
<th>$m/z$ 67, 81, 95</th>
<th>3-Vector Solutions</th>
<th>Vector-Matrix Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Number of Solutions in Family</td>
<td>$\Delta Q/Q_{exp}/(Q/Q_{exp})_{min}$</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>45</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>31</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>2</td>
<td>0.4%</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>3</td>
<td>0.8%</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>38</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>7</td>
<td>0.3–0.4%</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>4</td>
<td>0.6%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
<td>0.7%</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>XX</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>XX</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
matrix model is less than 0.8%, and in the fully constrained vector-matrix model ($\beta = 0$) is less than 0.05%.

The existence of multiple families for each number of factors complicates the choice of the best solution; now we must choose the best number of factors and the best family from that set of solutions. Criteria for choosing a family should be the same as for choosing the best number of factors: low $Q/Q_{exp}$, little structure in the residuals, and the factors must be physically meaningful. The families identify local minima in the $Q$ surface being explored during the iterative minimization. It would be tempting to choose the family with the most solutions; however, the number of solutions in one family may have a stronger relationship to the probability of entering a region of a local minimum in the $Q$ space and not have any intrinsic value. For example, during the iteration, the algorithm may enter a local minimum that has a large “opening” and therefore traps solutions that started from many seeds. Thus we reject a criterion for choosing a family based on the number of solutions in that family.

We now explore the solutions from the 3-vector model, choose the best solution, and identify the factors in this solution.

### 3.4.1 Results from the 3-vector model

The 3-vector model was solved by the PMF3 and ME-2 algorithms. The results from both algorithms were similar. Both found the same families at each number of factors in the solution. Since the results from the two algorithms are similar, and the vector-matrix cases are calculated with ME-2, we also present the ME-2 solutions of the 3-vector case. A comparison of
the results from the PMF3 and ME-2 algorithms when solving the 3-vector model is presented in Appendix H.3.

The simplest indication for choosing a good solution is a steep change of slope with the addition of each factor, but this criterion was not met for solutions of the 3-vector model. We therefore explore all solutions and use the criteria of finding physically meaningful factors to choose a good solution. Following the assumption that we will likely find the same factors from this dataset as Aiken et al. reported for the HR-MS data in that 2D factorization, we expect to need at least four factors in the solution. However, we also explore solutions with fewer factors in case new factors appear. The choice of the best solution of the 3-vector model is described Appendix F.1. The best solution of the 3-vector model has four factors, which are the factors from the HR-MS solution: OOA, HOA, BBOA, and LOA (Fig. 3.4). The features of these factors are discussed and compared to the factors from the vector-matrix model in Sect. 3.5.1 below.

3.4.2 Results from the vector-matrix model

Exploration of the vector-matrix model for this dataset is complicated by poor results in unconstrained solutions of the model. These poor results are manifested in non-physical factor mass spectra (Appendix H.4). To achieve factorization results with more physically meaningful mass spectra, we attempted two methods to constrain the mass spectra using a priori information. The first method was multiple linear regression, but the regression failed for most of the dataset (Appendix H.5). However, physically meaningful mass spectra were obtained by constraining the a priori spectra using the $\beta$ parameter (Sect. 3.3.4.1 above) within the vector-matrix model.
Fig. 3.4. The best 3-vector solution, which has four factors, solved with ME-2. The four factors are oxidized organic aerosol (OOA), hydrocarbonlike organic aerosol (HOA), biomass-burning organic aerosol (BBOA), and local organic aerosol (LOA). (a) Mass spectrum of each factor plotted vs. ion mass-to-charge ratio ($m/z$). Mass spectra are normalized to sum to 1. (b) Mass size distribution ($dM/d\log d_{va}$) plotted vs. particle vacuum-aerodynamic diameter ($d_{va}$) on a log scale. Size distributions are normalized so that the area under each curve sums to 1. (c) Mass contribution of each factor plotted vs. sampling date. The scale for LOA has been expanded to show the structure during low-concentration periods.
The use of the $\beta$ parameter with the vector-matrix model requires three additional choices: we must choose the number and source of the mass spectra to use as a priori information, and the value of $\beta$. We discuss the spectra first, and the choice of $\beta$ only for selected constrained spectra. We are likely to identify at least four factors (OOA, HOA, BBOA, and LOA) in the dataset based on the results of the 3-vector factorization. Since the 3-vector factorization had one solution with all four of these factors, we presume that all four factors should also be identifiable with the constrained vector-matrix model. Constraining between one and four factors gives fifteen possible combinations of a priori spectra, but exploring all of these combinations is beyond the scope of the present work. Because the unconstrained vector-matrix solutions produced non-physical factors, we have little confidence that constraining only one or two spectra will improve the solutions significantly. Thus we choose to constrain all four spectra.

Sources of the four a priori spectra include the HR-MS solution and the best solution of the 3-vector model, which had four factors. The two sources suggest in three combinations of a priori spectra: (1) all four spectra from the HR-MS solution, (2) all four from the 3-vector solution, or (3) some spectra from each of these solutions. We tried each of these three options. For the third, mixed-source option, we consider only the case taking OOA, HOA, and BBOA from the HR-MS solution and LOA from the 3-vector solution. We choose this combination since the OOA, HOA, and BBOA mass spectra from the 3-vector solution are very similar to those from the HR-MS solution, but the LOA mass spectrum from the 3-vector solution is less similar to the HR-MS LOA.
Fig. 3.5. Values of $Q/Q_{exp}$ for four-factor solutions of the vector-matrix model in which the factor mass spectra have been constrained to a priori spectra vs. $\beta$. The $\beta$ parameter can fully constrain the a priori spectra ($\beta=0$) or allow them to relax, as described by Eq. (3.5). The $Q/Q_{exp}$ values for solutions with $\beta \geq 0.3$ fall within the range of the unconstrained solutions (i.e., solved without a priori information about factor mass spectra) of the same model (grey region). The solutions in which the constrained OOA, HOA, and BBOA spectra come from the high-resolution MS (HR-MS) solution and the LOA spectrum comes from the four-factor solution of the 3-vector model (blue curve) is used for further analysis.
Comparison of the three combinations using fully constrained a priori spectra ($\beta = 0$) to the unconstrained vector-matrix solutions shows that the a priori spectra strongly influence the solution. This influence is shown by the constrained cases’ higher $Q/Q_{exp}$ values (Fig. 3.5). The increased $Q/Q_{exp}$ values suggest that the a priori spectra are not wholly consistent with the mass spectral structure measured in PToF mode. For example, constraining the mass spectra to the HR-MS spectra gives the highest $Q/Q_{exp}$ values of the constrained solutions (15% higher than the unconstrained vector-matrix solutions). The $Q/Q_{exp}$ value of this constrained solution is 5% higher than the best 3-vector solution. However, replacing the HR-MS LOA mass spectrum with the LOA spectrum from the 3-vector solution decreases the $Q/Q_{exp}$ value substantially, making it similar to the best 3-vector solution. In fact, even when the constrained mass spectra come from the best 3-vector solution, the $Q/Q_{exp}$ value is 2.5% higher than the unconstrained vector-matrix solutions with four factors. But, this constrained solution still has a lower $Q/Q_{exp}$ value than the solution from which the a priori spectra were obtained. Thus, the ability of the vector-matrix model to capture the real variability in the size distribution of each factor results in better fits to the data than the 3-vector model when using the same spectra. Of these three cases, the $Q/Q_{exp}$ values suggest that the spectra from the 3-vector model give the best fit.

In addition to the $Q/Q_{exp}$ values of these solutions, our choice for the best a priori factors is also based on the meaningfulness of the factor mass spectra. The HR-MS spectra provide stronger spectral information because they are derived from data with much higher SNR and high-resolution spectra, so we prefer to use these spectra to constrain the solution. However, the 3-vector LOA spectrum is more different from the HR-MS LOA. Although LOA is a small component, it is distinct enough to be resolved from the PToF data. Thus, the combined set of spectra with three factors from the HR-MS solution (OOA, HOA, BBOA) and the LOA
spectrum from the 3-vector solution is a good compromise, using mainly HR-MS spectral information and a better starting guess for the most disparate spectrum. We therefore choose the mixed source spectra as the best a priori information and explore solutions of the vector-matrix model in which we constrain these spectra.

### 3.4.2.1 Choosing a solution of the constrained vector-matrix model

To choose a solution of the constrained vector-matrix model, we must answer two further questions. First, might we be able to identify additional, physically meaningful factors by constraining the first four factors and then fitting more free factors? Second, to what degree is it appropriate to relax the constraint on the spectra, i.e., to increase $\beta$? We first choose the best number of factors while fully constraining four spectra, and then explore the relaxation of the constraint on the four a priori spectra in that solution.

The choice of the best solution of the constrained vector-matrix model is described Appendix F.2. The best solution of has the four a prior factors. In solutions with more than four factors, the additional factors are nonphysical splits of the HOA and BBOA factors. Therefore we continue to examine the four-factor solution and explore the effect of increasing to relax the constraint on the a priori spectra.

Two options for choosing an appropriate degree of relaxation ($\beta$) of the constrained factors are considered here. First, we can compare our factors to a priori information. This was the approach taken by Lanz et al. (2008) when they constrained one mass spectrum in their factorization of a 2D matrix. However, we have already imposed more a priori information on our solutions by constraining all of the mass spectra. Still, we could compare the time series of
our factors to the HR-MS time series or to external tracers. However, the external tracers were already used to support the choice of the HR-MS solution from which we have taken the a priori spectra, so this approach would be somewhat circular. If possible, we would prefer a more independent approach to choosing a solution for this dataset. An alternative option is to observe the fit residuals, which might indicate an appropriate degree of relaxation.

We examine two types of residuals metrics of the factorizations: the total residuals and the total $Q/Q_{exp}$ contribution summed across two dimensions of the 3D matrix, i.e., the total residual and total $Q/Q_{exp}$ as summed to form a time series, mass spectrum, or size distribution. The residual and $Q/Q_{exp}$ contributions summed to a time series are very similar across the range of $\beta$, and give no useful information about how to choose a solution. However, the residuals summed to form a mass spectrum show that many of the $m/z$’s $\leq 44$ have large negative residuals, i.e., the reconstruction assigns them more signal than is measured (Fig. 3.6a). But the large negative residuals of selected important $m/z$’s approach zero as the constraint is relaxed; i.e., $\beta$ is increased (Fig. 3.6b). In particular, two $m/z$’s show changes that are useful for choosing a solution. The residual of $m/z$ 44 is strongly negative in the fully constrained case, but becomes less negative as $\beta$ is increased to 0.06. As $\beta$ is increased further, the residual of $m/z$ 44 becomes more negative. Thus the inflection point at $\beta = 0.06$ marks the best fit for this important marker. In the same solution, the negative residuals for most other $m/z$’s tend toward zero values as $\beta$ is increased, and the high residual at $m/z$ 43 is reduced to approximately zero. Based on these trends, the solution at $\beta = 0.06$ may be the best solution of the constrained vector-matrix model.

The choice of the $\beta = 0.06$ solution is supported by the residuals and $Q/Q_{exp}$ contributions summed to a size distribution (Fig. 3.6c). The residuals summed as size distributions show that
Fig. 3.6. Residuals and $Q/Q_{exp}$ values for solutions of the constrained vector-matrix model with four factors.  

(a) Total residual (top) and $Q/Q_{exp}$ contribution (bottom) from each $m/z$. (b) Total residual for selected $m/z$’s as the solution is relaxed with increasing $\beta$. (c) Total residual (top) and $Q/Q_{exp}$ contribution (bottom) at each particle size. (d) Correlation between the a priori and calculated mass spectra as the constraint is relaxed with increasing $\beta$. The solution selected as “best” has $\beta = 0.06$ [black lines in (a) and (c) and red arrows in (b) and (d)].
solutions with tighter constraints have large negative residuals over the size range of particles with the greatest transmission in the AMS ($d_{wa} 50–700$ nm). However, the residuals across this size region are near zero in the solution with $\beta = 0.06$. As $\beta$ is increased further, the total residual becomes positive, indicating a worse fit in this region of the data. In addition, the $Q/Q_{exp}$ contribution vs. size does not change as $\beta$ is increased past 0.06. Thus the residuals and $Q/Q_{exp}$ contributions summed as size distributions confirm our choice of the solution with $\beta = 0.06$ as the best solution of the constrained vector-matrix model.

Finally, we revisit the option of comparing the factor mass spectra to the a priori spectra. The factor mass spectra have not changed dramatically; in fact, the correlations between the factor and a priori spectra in the solution with $\beta = 0.06$ are greater than 0.94 (Fig. 3.6d). We note that the LOA spectrum changes the least compared to the a priori spectrum from the 3-vector solution, showing that the a priori LOA spectrum was a better representation of the PToF data than the a priori spectra for the other factors. The HOA and BBOA spectra from the HR-MS solution change the most. This change is not surprising, based on the results of the fully constrained solutions with five factors, in which the fifth factor was HOA-like factor with the characteristic HOA peaks shifted to higher $m/z$’s. The HOA-like factor implies that the a priori spectra do not completely match the HOA spectrum for the PToF data. But relaxing the constraint in the four-factor solution allows a better fit of the higher $m/z$ peaks. However, the correlation between the solution spectra and the a priori spectra never show a dramatic change that might indicate that the solution has become so relaxed that the spectra are strongly distorted, and therefore do not suggest a particular value of $\beta$ as an appropriate relaxation of the solution.
4.2.2 Factors in the best solution of the constrained vector-matrix model

The factors from the best solution of the constrained vector-matrix model show the changing size and concentration of the particles in the four factors (Figs. 3.7–8, H.4). The factor

![Diagram](image)

**Fig. 3.7.** The best constrained vector-matrix solution, which has four factors. Four a priori mass spectra were provided as starting guesses: OOA, HOA, and BBOA from the HR-MS solution, and LOA from the best solution of the 3-vector model (Fig. 3.3). The a priori spectra were allowed to vary with $\beta = 0.06$, as described in Eq. (3.5). (a) Mass spectrum of each factor plotted vs. $m/z$. Mass spectra are normalized to sum to 1. (b) Mass size distribution ($dM/d\log d_{va}$), normalized so that each size distribution has unit area, plotted vs. $d_{va}$ on a log scale on the y-axis and vs. sampling date on the x-axis. The data have been binomially smoothed by one point each in time and size. Light-grey pixels have zero signal.
size-distribution–time-series matrices are normalized to unit area for each time step to highlight the shapes of the size distributions (Fig. 3.7). The contribution to the smallest and largest particle sizes, which have lower transmission into the AMS, is very noisy for all factors. In the middle size range, OOA and BBOA have somewhat narrower size distributions with smaller contributions to particles with \( d_{\text{wa}} < 100 \) nm and more mass in larger particles. In contrast, LOA appears in particle of all sizes, but most of its signal is between 80 and 600 nm. Finally, the HOA size distributions are also broad, and regularly include particles with diameters as small as 50 nm and larger than 700 nm. The HOA size distributions are less noisy than those of the other factors, especially at smaller sizes.

The size distributions of the factors do not usually change quickly, but an interesting exception occurs on 24 March, when a cold surge brought clean air to the Mexico City basin, decreasing the total submicron organic aerosol concentration to very low levels (Fig. H.5; Aiken et al., 2009; Molina et al., 2010). After midnight, the total organic concentration decreased to ~3 \( \mu \text{g/m}^3 \), the lowest concentration measured during the campaign. At 03:30 and 08:00, separate LOA plumes are measured. As the second LOA plume arrives, the HOA concentration begins to increase. An hour later, a dramatic increase in the concentration of OOA begins with the onset of photochemistry. As the OOA concentration increases, the mode of the HOA size distribution shifts from ~75 nm to ~300 nm over the course of three hours. This event is similar to a case during a 2003 campaign in Mexico City which has been studied in detail (Dzepina et al., 2009). However, in the 2003 case, the low initial background and lower wind speed and boundary layer allowed a better observation of the evolution of the emissions from the city. In contrast, other days have a larger effect of background concentrations and advection, and the dynamic changes
Fig. 3.8. Diurnal average size distributions of OOA, HOA, BBOA, and LOA from the best solution of the constrained vector-matrix model. (a) Image plots of diurnal average size distributions plotted (dM/d log dva) plotted vs. dva on a log scale on the y-axis and vs. hour of the day on the x-axis. The size distributions are not normalized to unit area. (b) Size distributions from selected hours plotted vs. dva on a log scale on the y-axis. These distributions are normalized to unit area so that the shapes of the distributions can be compared.
that are no doubt occurring in the aerosol species (e.g., Hodzic et al., 2010) are less clear at a
given fixed site.

HOA growth from smaller particles is not limited to this single event, but is also observed
in diurnal average size distributions (Fig. 3.8). HOA shows the greatest variation in size
distribution and concentration, with higher concentrations in the morning, and a shift to larger
sizes through the afternoon. Normalizing the size distributions to unit area shows more clearly
that HOA particles appear to grow to larger sizes throughout the morning (Fig. 3.8b). In
contrast, the BBOA and OOA size distributions show little variation. In general, OOA has
largest mode of all of the aerosol components, and higher concentrations later in the day,
whereas BBOA has higher concentrations in morning. BBOA particles transported from
regional sources may have undergone chemical processing and have probably already grown to
these larger sizes. Finally, LOA has the most dramatic changes in size distribution, but its
typically low concentrations make the averages noisy and it is difficult to determine whether the
average changes actually reflect particle growth.

3.5 Discussion

In this section we discuss three main points. First, we compare the best solutions of the 3-vector
and vector-matrix models to the HR-MS solution and a tracer method for estimating factor size
distributions. Second, we explain the insights gained from the size distributions of the factors
and the effect of some details of particle vaporization and bounce on PToF sampling and the
calculated factors. Finally, we discuss directions for future research on the application of 3D
factorization models to PToF and other datasets.
3.5.1 Evaluation of the assumptions of the 3-vector model

We compare the best solutions of the 3-vector and vector-matrix models to answer two related questions. First, is the 3-vector model appropriate for this dataset? Second, is one of these models better than the other for this dataset?

We first consider the appropriateness of the 3-vector model. Pere-Trepat et al. (2007) note that in their dataset, “it was found that there is sufficient size dependence in the [factor] compositions that the strict trilinear model does not hold.” However, these authors provide no details to explain how they determined that the 3-vector method was inappropriate for their data. We seek such indicators in the solution \( Q/Q_{exp} \) values and residuals. Overall, \( Q/Q_{exp} \) values show that the best solution of the constrained vector-matrix model fits the data better than the best 3-vector solution, although the differences are small (0.41 vs. 0.44, respectively; Fig. 3.5).

Examination of the residuals of these solutions shows that the constrained vector-matrix solution provides a better fit, but does not provide evidence that the assumptions of the 3-vector model fail for this dataset. The major difference in the residuals of the two solutions is seen by comparing the total residuals summed to form a size distribution. While the residual size distribution of the constrained vector-matrix solution is near zero or negative (Fig. 3.6), that of the 3-vector model is almost always positive, with larger residuals toward the larger sizes. The fact that the residual is always positive indicates that there are some real variations in the component size distributions that cannot be fit with the 3-vector model. However, \( Q/Q_{exp} \) summed to size distributions is similar between the two solutions. Thus the residuals are distributed differently in the solutions of the two models. Yet, we have found no distinct
Fig. 3.9. Comparison of the factors from the best solutions of the 3-vector and vector-matrix models and the HR-MS solution for (a) mass spectra, (b) average size distributions, and (c) time series. The grey shaded region in (b) denotes ± one standard deviation of the variation in time of the size distribution in the vector-matrix model.
characteristic of the residuals that indicates that one of the models is better or worse for this dataset.

We next compare the solutions from the 3-vector and vector-matrix models to understand the differences between these solutions. We first compare the mass spectra and time series from the 3D factorizations to each other and to the HR-MS solution. Then we will compare the size distributions using two methodologies.

First, we compare the factor mass spectra in Fig. 3.9a. The factor mass spectra of OOA, HOA, and BBOA are similar enough to each other to identify them as the same aerosol component by eye. The variation amongst the mass spectra of the same component from the HR-MS solution and the solutions of the 3-vector and vector-matrix models are within the observed variation for OOA, HOA, and BBOA factors in other urban datasets (Ng et al., 2011). However, the LOA mass spectrum shows a substantially different fraction of $m/z$ 58 between the 3D and HR-MS LOAs. Despite the high contribution of $m/z$ 58 to the LOA spectra, the 3D LOA spectra still retain the distinctive enhancement at $m/z$ 91 and, to a lesser extent, characteristic HOA peaks that dominate the shape of the HR-MS LOA spectrum. Overall, the factor mass spectra are similar between the two 3D factorization cases, and do not provide a clear reason to choose one solution over the other.

Second, we compare the factor time series in the 3-vector and vector-matrix solutions (Fig. 3.9c). The time series of OOA between the 3-vector and vector-matrix solutions, and of LOA between the same solutions, have very high correlations ($R > 0.96$). In contrast, the HOA time series and BBOA time series are less similar between the 3-vector and vector-matrix solutions ($R > 0.89$). For the HOA and BBOA time series, we observe several instances in
which mass is “traded” between these factors. We explain trading with an example. On 16 and 17 March, the 3-vector model attributes more mass to BBOA than does the vector-matrix model, but less mass to HOA than the vector-matrix model. Thus the HOA and BBOA factors have “traded mass” over this short time period. The trading of mass is the main difference between the HOA and BBOA time series for the two solutions, though the general trends are very similar in both models for the time series of both factors. HOA and BBOA have similar unit-resolution mass spectra, making them hard to separate in some datasets (Lanz et al., 2008); this similarity likely contributes to the trading in these factors’ time series. Aiken et al. (2009) reported that factoring high-resolution spectra substantially improved the separation of these two factors compared to factorization of the unit-mass-resolution spectral data. The improvement is the result of separating the information from HR ions at the same nominal m/z that have different contributions to HOA and BBOA. For example, HOA and BBOA have similar contributions from m/z 57 in the unit resolution mass spectra, but the HR-MS factors show that m/z 57 in HOA is almost exclusively from C₄H₉⁺, while C₃H₅O⁺ contributes a major fraction of the m/z 57 signal for BBOA. Nevertheless, the time series from our factorization of the unit-resolution PToF data with the vector-matrix model track the HR-MS time series more closely than do the time series from the 3-vector model.

We notice other interesting trends when comparing the time series from the 3D solutions to those of the HR-MS solution. First, both 3D models attribute more mass to BBOA than the HR-MS solution in the latter third of the campaign. Interestingly, this is the “low fire period” of the campaign, during which several non-AMS fire tracers and models indicate that fire activity was much reduced (Aiken et al., 2010). But comparison of the 3D and HR-MS solutions shows that this BBOA signal is trading with the OOA signal. Because we believe, based on the reasons
discussed above, that the HR-MS factorization of the high-resolution mass spectral data should
give more accurate results, the calculated BBOA during this period is likely an artifact of the
factorization. Second, the OOA time series from the 3D solutions are very noisy. However,
binomial smoothing of the 3D OOA time series by one point improves the correlation with the
HR-MS OOA from R = 0.68 to R = 0.81. The noise in the OOA factors from the 3D solutions is
likely caused by the low SNR of m/z 44 in PToF mode. Third, there is also only moderate
correlation between the 3D and HR-MS LOA time series (R= 0.66 and 0.62 for the 3-vector and
vector-matrix solutions, respectively). The 3D LOA factors are assigned less than half the
average mass contribution of the HR-MS solution (0.52, 0.65, and 1.39 µg/m³ for the 3-vector,
vector-matrix, and HR-MS solutions, respectively). However, the 3D solution mass spectra for
LOA have less contribution from the characteristic HOA peaks, and that mass may have been
traded to the HOA factor. Overall, the results from the 3D solutions are consistent with the
general characteristics of the HR-MS solution.

Third, we can make two comparisons of the size distributions from the two 3D
factorizations. First, we can compare the static size distributions from the 3-vector model with
the average size distributions from vector-matrix model. Second, we can compare the dynamic
diurnal average size distributions from the vector-matrix with estimated diurnal size distributions
from the 3-vector model obtained by scaling the fixed size distribution by the factor time series.
In addition, we also compare the dynamic size distributions to size distributions estimated from
tracer m/z’s (Zhang et al., 2005; Cubison et al., 2008; Wang et al., 2010). The size distributions
of m/z’s 44, 57, and 60 are used as tracers of OOA, HOA, and BBOA respectively. The tracer
size distribution is scaled by the slope of the MS-mode time series to the PToF time series for
that m/z, and then the estimation formula between the tracer and the component is applied. In
this work, we use tracer-component relationships derived from MS-mode data for OOA from Ng et al. (2011) and for HOA and BBOA from Aiken et al. (2010).

The comparison of the average size distributions shows some differences between the 3-vector and vector-matrix models (Fig. 3.9b). The size distributions from the two models are very similar for OOA and BBOA, consistent with the constancy of their normalized size distributions observed in Fig. 3.8. In contrast, the HOA and LOA size distributions show more differences between the two models. The HOA size distribution from the vector-matrix model is shifted to higher particle sizes compared to the 3-vector model. The LOA size distributions are even more different between the two models. The average LOA size distribution from the vector-matrix model is strongly shifted to larger particles compared to the distribution from the 3-vector model, with about a third less particle mass between 50 and 300 nm in the vector-matrix solution. However, the size distributions from the 3-vector solution are within one standard deviation of the average from the vector-matrix model for all factors. Thus it appears that the 3-vector model is not fitting variations in the real size distributions that can be captured by the vector-matrix model.

Now we compare the dynamic estimates of the size distributions from the 3-vector solution, vector-matrix solution, and tracer methods. Size distributions from selected daytime hours for the four factors are shown in Fig. 3.10 and are not scaled to unit area. Note that there is no tracer size distribution for LOA because this factor was not included in the tracer-based estimation method (Aiken et al., 2009).

The OOA size distributions are quite similar for the two 3D models. The OOA size distributions from the tracer method are also similar to those from the models, but are generally
Fig. 3.10. Comparison of average size distributions as calculated from the 3-vector model, vector-matrix model, and tracer method for OOA, HOA, BBOA, and LOA. The evolution of the size distribution of one aerosol type from 08:00–14:00 is shown in each column. The size distribution of each aerosol type at the same time is shown in each row. An estimate of LOA by the tracer method is not available.
narrower than the 3D model size distributions. In addition, the tracer method OOA size
distribution sometimes has negative concentrations for particles with diameters greater than 800
nm. Thus noise in the m/z 44 size distribution is carried through the tracer method, whereas this
noise is tempered by the contribution of other m/z’s present in the OOA mass spectrum in the 3D
models.

In contrast, the HOA and BBOA size distributions exhibit trading of mass between these
two factors in the 3-vector and vector-matrix models, as discussed above. HOA always has a
larger mass contribution and BBOA a smaller mass contribution in the vector-matrix model. The
vector-matrix model fits real variation in the size distributions and was determined in the
discussion above to be more consistent with the HR-MS solution, and thus more likely to be
correct. In addition to the differences in mass contribution, the modes of the size distributions
differ between the models. The size distributions for both HOA and BBOA are shifted to larger
particle diameters in the vector-matrix model, as discussed above. For HOA, the tracer size
distribution usually matches the vector-matrix distribution well, though the tracer distribution has
more mass at 08:00 than either of the 3D models. Like the tracer HOA size distribution, the
tracer BBOA size distribution generally follows the vector-matrix size distribution, but has a
slightly smaller mass concentration. The size distributions from the tracer method sometimes
show negative concentrations for large particles, but less severely than for OOA.

Finally, the LOA size distributions from the 3-vector and vector-matrix models agree
well at small particle sizes, but diverge for larger particle sizes. The vector-matrix model always
attributes more LOA mass to larger particles than does the 3-vector model. This trend is more
important in terms of the fraction of the mass in large particles as the day progresses.
In conclusion, it appears that both 3D methods can represent much of the real variation in the time series and mass spectra of the components. But the vector-matrix model appears to be superior for more accurately capturing the factor concentrations and for identifying real variations of the size distributions. While the 3-vector model captures the main modes in the size distributions reasonably well, it cannot identify particle growth events that are resolved by the vector-matrix method. However, the full advantage of the vector-matrix model may not be apparent with this dataset because it includes only a few strong periods of rapid changes in the size distributions. Only the OOA growth event on 24 March 2006 shows fast changes in the HOA size distribution (Fig. H.5). Furthermore, the appearance of LOA only in brief spikes makes it harder to understand this small component. Nevertheless, we prefer the vector-matrix model for exploring the size distributions of AMS components. However, the 3-vector model may be useful for simplified analysis and initial explorations of datasets.

### 3.5.2 Insights into ambient aerosol and PToF sampling

The size distributions of OOA, HOA, and BBOA are consistent with previous interpretations of these factors in Mexico City and other studies. OOA has the largest mode diameter of ~400 nm, and an asymmetric distribution with strong extension to smaller particles. This shape is consistent with the condensation of semivolatile gas-phase molecules onto smaller particles. BBOA has a slightly smaller mode at ~300 nm. Although BBOA is generated as primary aerosol from a combustion process, BBOA has a smaller fraction of ultrafine \((d_{\text{ua}} < 100\) nm) particles than OOA (McMeeking et al., 2005; Levin et al., 2010). In contrast, the size distribution of HOA is quite broad, with the largest fraction of ultrafine particles of any of these factors. The ultrafine fraction is consistent with the primary nature of this combustion source,
and with past measurements of freshly emitted vehicle-exhaust particles. Fresh exhaust particles have a mode at $d_{va} \sim 100$ nm, and a second mode with $d_{va} \sim 500$ nm is sometimes present (Canagaratna et al., 2004). Thus the broad HOA size distribution found here may not represent only fresh HOA. A similar, broad distribution of HOA particles was observed at an urban Pittsburgh sampling location (Zhang et al., 2005). The broad distributions could be explained by emission of many larger HOA particles and/or the growth of HOA particles as they are coated by condensation of secondary organic and inorganic species onto existing particle surface area during the day. For example, during the large OOA growth event on 24 March, the HOA size distribution grows as the OOA concentration increases (Sect. 3.4.2.2, Fig. H.5). During this event, it is unlikely that the HOA particles grow and remain as externally mixed HOA particles; rather, the HOA particles are most likely coated by OOA and ammonium nitrate, and thus have an HOA core and coating of secondary material. When these particles are sampled in the AMS, they have mixed HOA-OOA mass spectra and are recorded at the particle’s coated size. Thus the HOA size distribution grows to larger particle sizes, even though each particle has a mixed composition. Therefore, particle size distributions must be interpreted as showing the size distribution of particles containing an aerosol component, and not necessarily as the size distribution of externally mixed particles of any single component.

The size distribution of LOA gives us some new insight into the source of this component previously identified by Aiken et al. (2009). These authors identified this factor as local based on its spiky time series, which correlates with single-particle measurements of nitrogen-containing organic carbon particles and lead-zinc-containing particles that appear to come from local industrial sources (Moffet et al., 2008). The HR mass spectrum of LOA includes N-containing peaks characteristic of aliphatic amines ($m/z$’s 58 and 86) and stable $C_6H_5C_nH_{2n^+}$ ions.
characteristic of phenylalkyl compounds (m/z’s 91 and 105, McLafferty and Turecek, 1993).
The size distribution of LOA on the morning of 24 March, when LOA briefly has a large fraction
of the organic mass is bimodal with modes at $d_{wa}$~100 nm and ~270 nm. Since LOA appears to
come from an individual, local source, it may have a more variable size distribution than, e.g.,
HOA, which represents the average size distribution of millions of vehicles and other
combustion sources. Nitrogen-containing factors have been identified in some HR-AMS-PMF
analyses of urban datasets (Huffman et al., 2009; Sun et al., 2011) and their direct sources are
also uncertain.

The factorization of the PToF data may be influenced by the timescales of particle
vaporization in the AMS. A distribution of vaporization rates for the chemical constituents of
individual particles could contribute to the differences between the factor mass spectra from HR-
MS and the 3D factorizations of the PToF dataset. The aerosol components (HOA, BBOA,
OOA, or LOA) are not pure compounds, but rather mixtures of chemical compounds that may
not vaporize at the same rate. In PToF mode, fast-vaporizing compounds are measured at the
“correct” particle size. However, compounds that vaporize more slowly would be measured at a
later time, and thus at an apparently larger size, or even during a subsequent chopper cycle.
During this study, mass spectra were recorded in PToF mode every 50 µs. In comparison,
vaporization times at ~600 °C for NH₄NO₃ and (NH₄)₂SO₄ are 80 and 150 µs, respectively,
whereas polystyrene latex spheres (PSLs) have a vaporization time of 200 µs (Cross et al., 2009).
The PSL vaporization/ion flight time is longer because extra time is required to break the
polymer into smaller volatile molecules on the vaporizer. Some aerosol components can
vaporize even more slowly than PSLs. For example, PbCl is estimated to evaporate with a
timescale (1/e decay) of < 10⁵ µs (0.1 s), while other Pb compounds evaporate with a timescale
of $\sim 1.4 \times 10^8$ µs (2.4 min, Salcedo et al., 2010). Though the Pb cases are somewhat extreme, they demonstrate the possibility of very long vaporization times that can effect the measured PToF signal. Thus, PToF sampling is biased toward the faster-vaporizing compounds, and ions from slower-vaporizing compounds are more likely to be recorded as background. Furthermore, the slowly-vaporizing compounds likely produce a different distribution of ions than the fast-vaporizing components. Therefore the mass spectra obtained from factorization of the PToF data may better represent the fast-vaporizing compounds of each factor. In contrast, MS mode averages ions from compounds with different vaporization rates in many spectra, so ions from particles that evaporate in even a few seconds are still captured accurately. Thus, the factor mass spectra from the MS mode data should be a better representation of the average mass spectrum of the component.

Finally, we assess the calculated errors for the PToF data. The unweighted $Q/Q_{\text{exp}}$ values of the solutions near 1 (Fig. 3.3) suggest that the errors have about the right magnitude. However, $Q/Q_{\text{exp}}$ near 1 is not sufficient evidence that the error values are correct. Because 34% of the data matrix elements have values less than 0, which cannot be fit with the positivity constraint, these elements must contribute to $Q/Q_{\text{exp}}$. Even if every positive matrix element is fit within its estimated uncertainty and contributes $Q/Q_{\text{exp}}$ of $\sim 1$, the negative matrix elements must contribute additionally to $Q/Q_{\text{exp}}$. Thus our $Q/Q_{\text{exp}}$ values appear to be somewhat low, and the error estimates for the PToF data may be slightly overestimated.
3.5.3 Directions for future research

We first discuss the application of 3D factorizations to PToF datasets from ToF-AMS instruments and our recommendations for applying these techniques. Finally, we suggest potential applications for 3D factorization models to AMS instruments currently in development.

First, we reiterate that the analysis presented here has used the size distributions of unit-mass-resolution \( m/z \)'s. The unit-mass-resolution data in this study hindered separation of HOA, BBOA, and LOA because these factors have contributions from a common series of \( m/z \)'s. In contrast, the Aiken et al. (2009) study factored higher-resolution data, in which ions at each nominal \( m/z \) can be separated into the contrasting time series of individual ions. The time series of more ions might allow the separation of additional factors if the factorization were performed on the size distribution of high-resolution fragments. However, constructing the factorization matrix would require fitting high-resolution ions for each size-resolved mass spectrum, which is not yet part of the standard HR-AMS data analysis software (PIKA) and is a major project by itself. Nevertheless, this analysis might be possible with high-SNR datasets from locations with sufficient aerosol mass concentrations. Increasing the SNR of the PToF data by reducing the range of \( m/z \)'s sampled (DeCarlo et al., 2006) using an aerosol concentrator (Khlystov et al., 2005), or applying these techniques to C-ToF-AMS data (which has ~4 times more signal than the V-mode of the HR-ToF-AMS, DeCarlo et al., 2006) should also help improve factorization results.

The two vector-matrix models not used in the present study could be applied size-resolved chemical-composition datasets to explore different questions than we have considered here. The vector-matrix model in which the vector contains a factor size distribution and the
matrix shows how the chemical composition of that characteristic size distribution changes with
time (Fig. 3.1c) would likely identify modes of submicron aerosol and the sources and processes
affecting the those aerosol (Alfarra et al., 2004; Zhang et al., 2005). In contrast, the vector-
matrix model in which the vector contains a factor time series and the matrix shows the size
dependence of that factor’s chemical composition (Fig. 3.1d) assumes that particles from a single
source arrive at the receptor together and may have different chemical composition at different
sizes (Pere-Trepat et al., 2007). Application of these two vector-matrix models to size-resolved
chemical-composition datasets is of interest and would allow quantitative evaluation of their
appropriateness for describing these data.

Finally, we make two recommendations to researchers who wish to factor size-resolved
AMS datasets. First, we suggest beginning with 2D factorization of the high-resolution, mass
spectral mode data. In our case, the factors from the HR-MS solution were critical for
diagnosing the initial unsatisfactory solutions of the vector-matrix model. In addition, the 2D
factors were also useful for confirming when the 3D factors had split and were not physically
meaningful. Second, we recommend exploring the 3-vector model and at least one of the vector-
matrix models shown in Fig. 3.1. Comparison of the results of two models enables the
exploration of the appropriateness of each model’s assumptions.

In addition to their application to size-resolved chemical composition datasets, these 3D
factorization models have the potential to analyze the structure of other 3D datasets. The 3-
vector and vector-matrix models can be applied to any appropriate 3D data matrix for which the
model assumptions are appropriate, not just the specific data type described here. For example,
chemically-resolved thermal-desorption datasets are inherently three-dimensional. The 3D
factorizations described in this work could be applied to thermal denuder-AMS datasets (previously factored using 2-D methods, Huffman et al., 2009), or data from any of several thermal-desorption mass spectrometers, including the thermal-desorption particle-beam mass spectrometer (TDPBMS, Tobias et al., 2000), thermal-desorption chemical-ionization mass spectrometer (TD-CIMS, Smith and Rathbone, 2008), or the micro-orifice volatilization impactor coupled to a CIMS (MOVI-CIMS, Yatavelli and Thornton, 2010). As a second example, the thermal-desorption aerosol GC/MS-FID, or TAG, uses chromatography to separate organic compounds from thermally desorbed ambient aerosol, also forming an inherently 3D dataset (Williams et al., 2006; Williams et al., 2010). In the TAG dataset, the majority of the signal is present as an “unresolved complex mixture” which has not yet been analyzed in detail (Williams et al., 2010). In each of these cases, the third dimension of the data (thermal desorption temperature or chromatographic retention time) is expected to provide information distinct from the bulk mass spectra, and we expect that separation of additional factors should be possible from these datasets.

3.6 Chapter Conclusions

We have applied two 3D factorization models to three weeks of continuous HR-ToF-AMS size-resolved organic aerosol composition data from Mexico City. Preparation of the data for factorization required a method for estimating the precision of the measured data, developed here for the first time. In the 3-vector model, each factor is composed of a characteristic chemical composition (mass spectrum), a characteristic size distribution, and the time series of the mass concentration of that component. In this model, the mass spectrum and size distribution are constant over the course of the measurements. In contrast, in the vector-matrix model in
which the vector is a mass spectrum, the matrix shows the changing size distribution of that chemical component with time. The vector-matrix model has more degrees of freedom than the 3-vector model; the additional freedom provides greater ability to fit the dynamic nature of the data, but also to be distorted by noise in the dataset. Noise hampered initial results of the vector-matrix model, but physically meaningful factors were obtained after partially constraining the mass spectra using a priori information and a new constraining metric. For this dataset, four factors were identified that were consistent with factors obtained by Aiken et al. (2009) from factorization of the bulk (i.e., not size-resolved) HR-MS organic measurements from the same instrument. These factors represent oxidized organic aerosol (OOA), hydrocarbon-like organic aerosol (HOA), biomass-burning organic aerosol (BBOA) and a locally emitted organic aerosol (LOA). However, the mass spectra of these factors are not identical to those from the HR-MS solution. These differences may be due to noise and/or a fraction of slowly-vaporizing compounds whose ions are averaged into the total signal in MS mode but are recorded as part of the background in PToF mode.

The results of the vector-matrix model show diurnal cycles in the size distribution of HOA and suggest growth by condensation of secondary species onto pre-existing HOA particles, especially during an OOA growth event on 24 March 2006. The size distributions of HOA and BBOA are consistent with source size distributions. In addition, these size distributions are less noisy and likely more robust than those obtained previously by tracer methods, and could be used for future cloud condensation nuclei (CCN) and hygroscopicity studies (Cubison et al., 2008; Gunthe et al., 2009; Wang et al., 2010).
The vector-matrix model appears to capture real variability in the size distributions that cannot be captured in the 3-vector model. While the 3-vector model captures the main modes in the size distributions reasonably well, it cannot identify particle growth events that are resolved by the vector-matrix method. We suggest that others who apply this technique first factor a 2D version of the HR-MS data to understand the main trends in the dataset and as a basis for understanding factors from 3D model solutions.

Other versions of the vector-matrix model are possible and could be applied to this type of dataset to explore other questions about aerosol evolution. Finally, these techniques can be applied to other 3D datasets, especially those obtained by measuring thermal desorption aerosol mass spectra or chromatographically resolved aerosol composition.
Chapter 4

Conclusions

4.1 Thesis Summary

The work for this thesis has three main parts. First, a free, open-source software package was developed for the fast visualization of factor analytical results (Chapter 2, Appendix A). Second, an organic mass spectral dataset from Pittsburgh in 2002 was analyzed by a bilinear unmixing model, solved using an algorithm which positively constrains the factor elements to be positive (Positive Matrix Factorization, PMF; Chapter 2). The behavior of PMF solutions with AMS data was characterized using several 2- and 3-factor synthetic datasets with realistic noise (Chapter 2). Third, a size-resolved organic mass spectral dataset from Mexico City in 2006 was analyzed by two three-dimensional (3-D) factorization models to study variations in the size distributions of organic aerosol components (Chapter 3).

The software package developed in this work is known as the PMF Evaluation Tool, or PET. The PET executes PMF runs for a range of two parameters: number of factors in the solution, p, and rotational parameter, FPEAK, or random start generator, seed. The PMF output from all p-FPEAK or p-seed combinations is saved, and the PET calculates the solution residuals and various statistics of the fit. The main panel of the PET displays the factors and residuals for one solution, and the user can easily view different solutions or overlay the results of multiple solutions. An additional panel in the PET displays correlations between factor and reference mass spectra and time series. The PET has been coded in Igor Pro (WaveMetrics, Portland, OR).

To characterize the behavior of PMF with AMS datasets, 2- and 3-factor synthetic datasets were constructed based on earlier factorization of an AMS dataset using a custom algorithm. When the synthetic datasets were solved with “too many” factors (i.e., with more factors than were used to create the synthetic dataset), the solution factors included mixes of the original factors, or “splits” of the original factors into multiple factors with similar mass spectra and time series. The mass spectra of the mixed factors had high correlations with mass spectra from the AMS Spectral Database. The high correlation between the mixed factor spectra, which cannot be real for the synthetic dataset, and real AMS mass spectra demonstrates that mass spectral correlations are not sufficient for determining whether factors from PMF solutions represent real aerosol components. The behavior of split factors suggests that solutions that include factors with highly similar mass spectra and/or time series be viewed with caution to determine whether the splitting has chemical meaning or is a mathematical artifact.

The factorization of the Pittsburgh dataset by PMF led to the identification of three factors. Hydrocarbon-like organic aerosol (HOA) is produced mainly from combustion sources such as vehicle emissions, and may sometimes include cooking emissions. The time series of HOA usually correlates well with other combustion tracers such as NOx, CO, and black carbon. In addition to HOA, two oxygenated organic aerosol (OOA) factors were identified. OOA is predominantly secondary organic aerosol (SOA) that is found ubiquitously in datasets from around the world. The two types of OOA have different volatility characteristics and represent endpoints of SOA processing in the atmosphere. Semi-volatile OOA (SV-OOA) represents “fresh” OOA and has a mass spectrum that includes both hydrocarbon-like and oxygenated ion
fragments. The time series of SV-OOA correlates well with particulate ammonium nitrate, a semi-volatile inorganic aerosol component. The more highly oxidized low-volatility OOA (LV-OOA) represents “aged” OOA and has a more oxidized mass spectrum. The time series of LV-OOA correlates well with $O_x (NO_2 + O_3)$, which is the sum of odd-oxygen gas-phase oxidants, and to ammonium sulfate, a secondary, low-volatility inorganic aerosol component.

Two three-dimensional factorization models were applied to a size-resolved organic aerosol composition dataset from Mexico City. In the 3-vector model, each factor is composed of three vectors that represent a factor’s chemical composition, size distribution, and mass concentration in each time step. In this model, neither the aerosol composition nor size distribution can change with time. In the vector-matrix model, each factor is composed of a vector and a matrix. In the form used in this study, the vector represents a factor’s chemical composition and the matrix represents the time-varying size distribution of that factor. The vector-matrix model has more degrees of freedom than the 3-vector model; the additional freedom provides greater ability to fit the dynamic nature of the data, but also to be distorted by noise in the dataset. Noise hampered initial results of the vector-matrix model, but physically meaningful factors were obtained after partially constraining the mass spectra using a priori information and a new constraining metric. The vector-matrix model appears to capture real variability in the size distributions that cannot be captured in the 3-vector model. While the 3-vector model captures the main modes in the size distributions reasonably well, it cannot identify particle growth events that are resolved by the vector-matrix method.

The best solution of each model had four factors: HOA, OOA, biomass burning organic aerosol (BBOA, whose mass spectrum is distinguished from HOA by the enhancement of signal at $m/z$’s 60 and 73 in BBOA), and a nitrogen-rich, locally occurring organic aerosol (LOA). The
results of the vector-matrix model show diurnal cycles in the size distribution of HOA and suggest growth by condensation of secondary species onto pre-existing HOA particles, especially during an OOA growth event on 24 March 2006. The size distributions of HOA and BBOA are consistent with source size distributions. In addition, these size distributions are less noisy and likely more robust than those obtained previously by tracer methods.

4.2 Synthesis of Findings

Factor analytical techniques are useful for extracting components from Aerodyne aerosol mass spectrometer (AMS) organic aerosol datasets. Factor analytical techniques are ideal for AMS datasets because the ionization technique (electron-impact ionization, 70eV) gives rise to repeatable fragmentation from vaporized molecules, creating inherent internal correlations in the dataset. In particular, Positive Matrix Factorization (PMF for 2-dimensional factorization) and the Multilinear Engine (ME-2 for a variety of models) are most capable of providing physically-meaningful factors because the both the signals in the chemical profiles and the mass concentrations in the time series are constrained to be positive. Another advantage of PMF is that it solves the fit by minimizing a weighted $\chi^2$ term, and so data with strong information content (i.e., high signal-to-noise-ratio, SNR) can have a strong contribution to the fit, while the influence of data with little information content (i.e., low SNR) can be decreased. Consequently, good estimates of the weights (precision of the measurement) are required for successful application of PMF.

The study of synthetic datasets is a useful exercise for characterizing new factor analytical models and algorithms. While it is not possible to construct every possible synthetic dataset, exploration of representative datasets can give good insight into the behavioral
relationships between the data and the model. When the synthetic datasets will be factored by weighted-minimum algorithms such as PMF and ME-2, the construction of estimates of the “precision” for these data is also necessary. These estimates should be constructed in a manner as close to the real data as possible so that the behavior results from the synthetic dataset may be generalized to real datasets.

The application of weighted factor analytical algorithms to real datasets is somewhat tricky. First, a good understanding of the measurement precision is imperative. In addition to considering the uncertainty of the measured signal at a detector, calculating the “chemical signal” for complex datasets usually involves multiple steps, such as baseline estimation, background subtraction, peak fitting, and smoothing. The uncertainty for each of these steps must be propagated with the measurement uncertainty. While the weighting in algorithms such as PMF and ME-2 can protect the analysis from low-SNR data, it is best to obtain high-SNR data in the first place when possible. Second, the modeler must make a subjective decision to choose the number of factors in a factor analytical solution. This decision should be supported by evidence external to the factor analysis, i.e., the time series of species that were not included in the factorization matrix. For AMS data, correlations with the time series of external time series are a stronger metric than correlations with external mass spectra, which tend to contain many of the same fragments because of the patterns generated by electron impact ionization. Third, current factorization models do not match our understanding of the atmosphere. For example, two-dimensional models assume that the chemical composition of the aerosol is unchanging, but we know that atmospheric reactions change the aerosol composition. Similarly, the three-dimensional models used in this work assume that at least one dimension of the aerosol (e.g., chemical composition or size distribution) is static, but we know that the system is dynamic.
Thus we should expect that the factorization models will not fit the data perfectly, and that the discrepancy between the conceptual and mathematical models will be reflected in the factors and/or the residuals of the factorization solutions. Researchers should explore these effects in their dataset to understand the limitations of the factorization model and the consequent interpretation of the factors.

In most two-dimensional AMS datasets, we can reliably separate HOA, OOA (either as one component or both SV-OOA and LV-OOA), and when it is present in sufficient quantities, BBOA. Three-dimensional models open avenues to exploring more complex datasets. The factorization of a size-resolved organic aerosol composition dataset demonstrates these possibilities. We found the evolving size distributions of HOA, BBOA, OOA, and a nitrogen-rich, locally occurring organic aerosol. The vector-matrix model is able to capture real variability in the aerosol size distributions that cannot be captured in the vector-matrix model. The factor size distributions are less noisy than those previously estimated by a tracer method and may be useful for cloud-condensation nuclei (CCN) and hygroscopicity studies. While the inclusion of the aerosol size information in the 3-D factorization did not help identify additional factors in this dataset, application of the same methods to a size-resolved aerosol composition dataset with higher SNR may have different results.

4.3 Directions for Future Research

4.3.1 Better Understanding of “Split” Factors

One of the main indicators that a solution may have “too many” factors is that the factors “split” into multiple factors. These split factors usually divide the mass attributed to a single factor (in a solution with fewer factors) into two or more factors (in a solution with a greater
number of factors); the split factors may have similar mass spectra. The factorization of some
AMS datasets has shown that after some factors have split, solutions with even more factors may
distinguish new, physically meaningful factors that were not separated in solutions with fewer
factors. A solution acceptable for interpreting and presenting the results can be obtained by
adding the split factors back together to reconstruct the main factor.

However, the fact that factors split must have some additional meaning; this has not yet been
explored in great detail. The occurrence of split factors indicates a chemical difference between
these factors. There are three likely causes for chemical changes in a factor: instrumental
variation, inadequate estimation of the precision of the measurements, or atmospherically-
relevant variation in aerosol composition. Instrumental variations could be caused by changing
the instrument tuning or changing the vaporizer temperature in such a way that the mass spectra
change during a portion of a measurement campaign. In contrast, if the measurement precision
has been estimated poorly, some less important information may have too much weight and
cause the algorithm to try to fit it. But we are most interested in atmospherically-relevant
variation in aerosol composition. Thus it is important to distinguish between factors that split
because the weights have been applied incorrectly and those that may represent real variation in
the chemical composition of the aerosol. Identification of changes in OOA composition are
most likely to be supported by measurements of more tracer species, especially of semi-volatile
organic compounds (SVOCs) and intermediate-volatility organic compounds (IVOCs) as
techniques to measure these challenging compounds become available. Better understanding of
split HOA factors may come from more measurements of HOA sources and the variation in their
composition.
4.3.2 Application of Factorization to New Datasets

New understanding of aerosols is likely to come from the application of factor analytical techniques to other complex datasets that incorporate different types of multiplexing. Such data can be acquired by two types of existing instruments. One class of instruments measures thermally-desorbed, chemically resolved aerosol composition; these datasets have dimensions of time, chemical composition, and thermal desorption temperature and separate aerosol components based on volatility. The second class of instruments measure chromatographically-separated, chemically resolved aerosol composition; these datasets have dimensions of time, chemical composition and chromatographic retention time and separate aerosol components based on volatility or molecular polarity. The third dimension in each of these cases is expected to provide information distinct from the bulk mass spectral information, and we expect that separation of additional factors should be possible from these datasets.

A second interesting avenue for factor analysis is its application to aerosol composition datasets acquired with soft ionization techniques. However, these datasets must conform (sufficiently) to the assumptions of the model: that each sample is the linear sum of aerosol of different compositions. Thus these datasets must be quantitative, or at least the ionization must be repeatable (i.e., without biases, or the biases must be constant), and good estimates of the measurement precision must be made. The best soft-ionization techniques for generating datasets for factor analysis would have two qualities. First, the ionization would not be very selective, and therefore be able to ionize a wide variety of compounds. Second, the ionization would be soft enough to have little fragmentation (or less than electron impact) so that the mass spectra have more signal from molecular ions and/or larger fragments. Such datasets should
have different internal patterns than electron impact datasets and hopefully help identify more molecular characteristics of already familiar factors such as HOA and OOA.

### 4.3.3 Creation of New Factorization Models

ME-2 is a powerful and flexible tool for factorization of multidimensional datasets. This tool is designed to be flexible to solve factorization models that can be expressed as the sum of products of factor elements. In addition, the modeler can specify constraints on the factor elements, or on the relationship of factor elements to each other. Thus we have a tool to solve a multitude of models that we create. However, such models must be carefully constructed and validated, which is a challenge on its own.

It would be beneficial to apply and understand the vector-matrix models that were not applied in this work. That is, it would be useful to apply them to size-resolved aerosol composition data and understand whether they produce factors that give additional insight into aerosol sources or processes. Other models already exist that may be useful for three-dimensional aerosol datasets, including Tucker2 and Tucker3 models (Tucker, 1966). These models have been applied in other areas of environmental factor analysis, but to our knowledge, not to aerosol composition data.

One area of interest that is unsatisfied by existing models is the changing aerosol composition in two-dimensional datasets. Current two-dimensional models assume that aerosol composition cannot change; however we know that this does not accurately describe the atmosphere. Development of such a model should be possible using constraints in ME-2. However, but such a model will require careful validation of the parameterization for allowing
aerosol composition to change, and testing will require suitable datasets that can be used to test the parameterization assumptions.

4.3.4 Maintenance of Good Software Tools

This work performed in this thesis has been facilitated by having a robust and fast visualization tool. The availability of this or similar tools will be essential for moving forward in this field. The existing tool is open source and can be modified for new data and models.
References


Canagaratna, M., Worsnop, D., Zhang, Q., and Jimenez, J. L.: Aerosol Mass Spectrometry of
Organic Aerosol in Houston, Texas, International Aerosol Conference, St. Paul, MN, 2006,
Canagaratna, M. R., Jayne, J. T., Jimenez, J. L., Allan, J. D., Alfarra, M. R., Zhang, Q., Onasch,
T. B., Drewnick, F., Coe, H., Middlebrook, A., Delia, A., Williams, L. R., Trimborn, A.

Cass, G. R.: Methods for Sulfate Air Quality Management with Applications to Los Angelos,

Chan, T. W., and Mozurkewich, M.: Application of Absolute Principal Component Analysis to
Size Distribution Data: Identification of Particle Origins, Atmos. Chem. Phys., 7,

Distributions Using Absolute Principal Component Analysis, Atmos. Chem. Phys., 7,

Chow, J. C., and Watson, J. G.: Review of PM2.5 and PM10 Apportionment for Fossil Fuel
Combustion and Other Sources by the Chemical Mass Balance Receptor Model, Energ.

Christensen, W., Williams, B., and Reese, C.: Bayesian Approaches for Pollution Source
Location Identification and Apportionment, American Association of Aerosol

Costabile, F., Birmili, W., Klose, S., Tuch, T., Wehner, B., Wiedensohler, A., Franck, U., König,
K., and Sonntag, A.: Spatio-Temporal Variability and Principal Components of the
Particle Number Size Distribution in an Urban Atmosphere, Atmos. Chem. Phys., 9,


Crosier, J., Jimenez, J. L., Allan, J. D., Bower, K. N., Williams, P. I., Alfarra, M. R.,

Cross, E. S., Onasch, T. B., Canagaratna, M., Jayne, J. T., Kimmel, J., Yu, X. Y., Alexander, M.
L., Worsnop, D. R., and Davidovits, P.: Single Particle Characterization Using a Light
Scattering Module Coupled to a Time-of-Flight Aerosol Mass Spectrometer, Atmos.

Cubison, M. J., Ervens, B., Feingold, G., Docherty, K. S., Ulbrich, I. M., Shields, L., Prather, K.,
Hering, S., and Jimenez, J. L.: The Influence of Chemical Composition and Mixing State
of Los Angeles Urban Aerosol on CCN Number and Cloud Properties, Atmos. Chem.


Fuzzi, S., Decesari, S., Facchini, M. C., Matta, E., Mircea, M., and Tagliavini, E.: A Simplified


Paatero, P.: Interactive Comment on “Interpretation of Organic Components from Positive Matrix Factorization of Aerosol Mass Spectrometric Data” by I. M. Ulbrich Et Al.,


Thurston, G. D., and Spengler, J. D.: A Quantitative Assessment of Source Contributions to


Weber, R. J., Sullivan, A. P., Peltier, R. E., Russell, A., Yan, B., Zheng, M., de Gouw, J.,


Zhang, Q., Alfarras, M. R., Worsnop, D. R., Allan, J. D., Coe, H., Canagaratna, M. R., and


Appendix A

Explanation of the PMF Model and Algorithm

1. Introduction

In chemical systems with linear, additive properties, observing a mixture of compounds shows a linear combination of the additive property, which can be modeled as a system of linear equations. Spectroscopy of radiation-absorbing gases is a common example of this phenomenon. Each gas has a unique pattern of absorption across a set of wavelengths, and the absorption of a mixture of gases (for example, a column of the earth’s atmosphere viewed from space to the ground) is the sum of the absorption by each gas, as long as the total absorption is small (less than about 10%).

Consider ambient particles in the lower altitudes of Earth’s atmosphere. Particles are considered “primary” when they are emitted directly from a source or a mechanical process. Examples of primary particles include soot from cars, power plants, meat cooking, etc., salt particles formed from ocean spray, and dust from road and desserts. Particles are considered “secondary” when they form from collisions of gas molecules in the air. This has been observed when ammonia molecules (primarily released from animal waste and fertilizer) meet nitric acid (formed from photooxidation of nitrogen oxides emitted from combustion, such as in cars and power plants) and react to form solid particles. It is also known that gaseous organic carbon compounds can react in the atmosphere to form low volatility species that condense to form liquid or solid particles (Finlayson-Pitts and Pitts, 2000).
Particles in the atmosphere can be sampled and their chemical components analyzed. Such analyses are usually performed on a batch of particles collected over a time period that may range from a few seconds to an entire day. For species that do not react in the atmosphere, the amount of each species measured will be a linear combination of the contributions of the various sources of those compounds. A set of measurements from many samples (time steps) then forms a matrix in which each row represents the chemical composition of one sample and each column represents the concentrations of one chemical species measured over the time series of measurements.

With the knowledge that the matrix of observations is a linear combination of some number of sources with respective mass contributions, we can attempt to decompose the observations into their contributing components as matrix product, where one matrix gives the chemical composition of a source (known as a “source profile”) and the other their contributions to the particle mass over time (often referred to as a “source strength”), or expressed mathematically,

\[ X_{mn} = G_{mp}F_{pn} + E_{mn} \]  

(A1)

where \( X \) is the \( m \times n \) matrix of data points,

\( G \) is the \( m \times p \) matrix of strengths (time series) of \( p \) sources,

\( F \) is the \( p \times n \) matrix of the profiles of \( p \) sources,

and \( E \) is the \( m \times n \) matrix of errors or residuals of each data point not fit by \( p \) sources.

The matrix \( X \) is often oriented such that \( m \) (rows) is the number of measurements over time and \( n \) (columns) is the number of chemical components considered. Note that
if $X$ is transposed, then the $G$, $F$, and $E$ matrices will also be transposed and the data matrix is reproduced by $X^T = F^T G^T + E^T$.

A number of so-called “source apportionment” techniques have been utilized by researchers in these fields. Three techniques will be discussed here briefly, and the remainder of the paper focuses on the last technique.

1.1 Chemical Mass Balance Model (CMB)

The concept of a mass balance of particle components was conceived and applied in the 1970s and was the first source apportionment technique applied to ambient particles (Henry, 1997). The premise of the chemical mass balance model is that the data matrix $X$ can be reconstructed through a linear combination of known profiles of the sources (i.e., the matrix $F$). The model then solves the systems

$$x_i^T = g_i^T F_{pm} + e_i^T$$

(A2)

where $x_i^T$, $g_i^T$, and $e_i^T$ are the $i^{th}$ row of the data matrix $X$, the strength matrix $G$, and the error matrix $E$, respectively,

and $F$ is the matrix of known source profiles.

This is a simple case of least squares fitting (resembling the familiar $Ax=b$ system after taking the transpose) for each row of the data matrix $X$. While the method seems simple and convenient mathematically, considerable effort is required to acquire the source profiles for all sources that may contribute to the ambient particles.

The method is susceptible to several failure mechanisms. First, the number of chemical components must equal or exceed the number of sources (Coulter, 2004); that
is, the size of $F$ must have $n \geq p$. If we presume that the source profiles are linearly independent, then the rank of $F_{nxp}^T$ is $p$, ensuring that a least-squares solution exists.

Second, source profiles should be obtained for all potential sources; if sources are omitted, their contributions will be incorrectly apportioned to other sources. The source profiles must be collected and analyzed in the same manner as the dataset to be evaluated; this restricts the ability of researchers to share data collected with different methodologies. These first two constraints require the experimenter to obtain sufficient chemical characterization of the particles to maintain the proper dimension of the source profile matrix. If too many sources are considered, $p > n$ and additional laboratory work is required! The third caution is that the method is robust only if the known source profiles are very similar to the particles emitted by the source during the ambient study. If some component differs between the measured profile and the particles emitted by that type of source at a later time (e.g., sulfur was removed from a fuel or an engine was sampled at a non-standard power setting), the least squares fit will be inappropriate. Finally, some secondary particles do not have known composition,\(^1\) so their profile cannot be included in $F$ and these particles will not be properly apportioned. Recent studies suggest that in some size ranges, approximately 66% of particle mass may be secondary particles (Zhang et al., 2005).

### 1.2 UNMIX

Henry (2003) developed UNMIX as a method utilizing singular value decomposition (SVD) to find edges in a space of N-source dimensions. UNMIX uses a

---

\(^1\) Secondary particles may comprise only a few molecules at the instant of their formation and are extremely difficult to capture for chemical analysis. In the case of secondary particles formed only from organic compounds, many compounds could be involved and the number of combinations enormous.
custom algorithm to determine the number of sources with significant contributions to the data and reduces the SVD matrices to this size. Data are projected onto an (N-1)-dimensional plane and the contribution of each source is determined by fitting points as edges and vertices of the space (Henry, 2003). UNMIX only returns unique results, not allowing the user to explore rotations. Henry does point out that his philosophy may be uncommon, stating in the user’s manual for EPA’s version of the software (Henry, 2000)²

The goal is to let the data speak for itself…. Source compositions and contributions must be non-negative. Unfortunately, it has been shown that non-negativity conditions alone are not sufficient to give a unique solution, more constraints are needed (Henry, 1987). Under certain rather mild conditions, the data itself can provide the needed constraints (Henry, 1997). This is how UNMIX works. However, sometimes the data do not support a solution. In this case UNMIX will not find one. While some might judge this a disadvantage, it is actually a positive benefit to the user. Few modeling approaches let the user know clearly when a reliable solution is not possible. No solution is better, if not more satisfying, than an unreliable one.

1.3 Positive Matrix Factorization (PMF)

The rest of this paper will review Paatero and Tapper’s (1993, 1994; Paatero, 1997a) variant of factor analysis and principal component analysis known as Positive Matrix Factorization (PMF), focusing on Paatero, 1997a. While other methods of principal component analysis modify the data matrix by taking logarithms, centering rows or columns by their mean values, scaling by row or column norm, or scaling by a correspondence matrix, Paatero and Tapper claim (1993) to make an optimal scaling of the data matrix by dividing each data element by its estimated error (standard deviation). They argue that this scaling has the best physical relationship between the data and its

² Citations in this paragraph were made in the original text.
measurement technique. The system described in Equation A1 is solved by an iterative technique which minimizes the weighted norm

$$Q = \left\| \frac{(X - GF)}{\sigma} \right\|_F^2$$

(A3)

where $X$ is the $m \times n$ matrix of data points,

$G$ is the $m \times p$ matrix of strengths of $p$ sources,

$F$ is the $p \times n$ matrix of the profiles of $p$ sources,

$\sigma$ is the matrix of standard deviations of the errors of the elements of $X$,

$E$ is the $m \times n$ matrix of errors or residuals of each data point not fit by $p$ sources,

$\div$ represents element-by-element division, and

$\|B\|_F$ represents the Frobenius Norm, $= \sqrt{\sum_i \sum_j B_{ij}^2}$.

The technique is known as positive matrix factorization because both $G$ and $F$ are constrained to have non-negative values, representing the non-negative nature of chemical properties in the score matrix and their non-negative contributions in the loading matrix.

As in UNMIX, the number of sources, $p$, is unknown. The correct value of $p$ in PMF is usually determined by plotting many values of $p$ against their respective minimized $Q$ and choosing the $p$ where the line changes slope. Identification of these $p$ sources is performed by the analyst and must be carefully defended in publications.

PMF has been utilized extensively in the literature, with reference (Paatero, 1997a) being cited 87 times at this writing; 18 of these citations are from articles published in 2005. In addition to applications for determining the sources of particles in all regions of the globe, PMF has been used to investigate the sources of volatile organic compounds (Jorquera and Rappengluck, 2004; Zhao et al., 2004), ozone levels (Rizzo
and Scheff, 2004; Wang et al., 2003), endocrine-disrupting contaminants in coastal sediments (Pere-Trepat et al., 2004), and other subjects in which positivity constraints are required by the nature of the data.

The remainder of this work explains the mathematics and algorithm of PMF and makes use of that author’s algorithm with a simple, synthetic data example to explore the ability of PMF to reproduce a known linear combination. The final section describes the application of PMF to my own research and asks questions about PMF’s capabilities in that application. I hope to resolve these questions during the course of my doctoral research.

2. Mathematical Formulation of Positive Matrix Formulation

This section describes the method of least squares with non-negative factors described by Paatero in reference (Paatero, 1997a). This review focuses on the sections of Paatero’s article that explain the least-squares solution of the model, with many references to his earlier papers (Paatero and Tapper, 1993, 1994) which support some claims in the current paper.

2.1 Paatero’s Introduction

Factor analysis (FA) and principle component analysis (PCA) are techniques commonly used in physical and social sciences to extract additional information from data. In order to better fit the data, some form of normalization is usually utilized. Standard normalization techniques include “centering” variables by subtracting the column mean, taking logarithms of data spanning several orders of magnitude, or
dividing by row norm, column norm, or the correspondence of variables in a column. Paatero and Tapper (1993) argue that none of these scalings have a suitable connection to physical or chemical data, and that they ignore the best connection to the actual measurement – its estimated error (standard deviation). While the scalings used in FA and PCA are limited to treating entire rows or columns, normalizing by each measurement’s standard deviation allows scaling of individual data elements.

To show that their scaling is “optimal,” Paatero and Tapper utilized a well-known synthetic (i.e., constructed) dataset “QUAIL ROOST II” developed in the early 1990’s for testing the abilities of source apportionment models (Paatero and Tapper, 1993). Paatero and Tapper analyzed the dataset using many normalization schemes and showed that dividing by the measurement standard deviation was indeed “optimal” because it gave the fit with the smallest least-squares error (Figure A.1). This certainly demonstrates that the scaling is an improvement over other methods, but may not actually prove that it is the best scaling possible.

In addition to providing this “optimal” scaling of the data, dividing each point in the data matrix $X$ by its measurement standard deviation improves the ability of the method to handle three important cases which are common in environmental data. First, one column of the matrix (the concentrations of one compound) may have low signal to noise. For this column, then, the weighted $X_{ij}/\sigma_{ij}$ is less than 1, and the column has a lessened contribution to the matrix factorization. When this situation is encountered in PCA, there may be one factor which consists almost entirely of this column and represents the noise in this column. The user must identify this and realize that this is not a factor of importance and exclude or ignore it later. The “optimal” scaling prevents this
Figure A.1. Fit of QUAIL ROOST II data by several normalization schemes. Quality of the least squares fit (Chi² Value, i.e., Q) plotted against number of components (p). The symbols represent scalings as follows: □ No Scaling △ Correspondence Analysis ○ Column Norm ● Row Norm ▲ Balanced Scaling * Optimal Scaling

The balanced scaling (an iterative variant of weighted singular value decomposition) and optimal scaling show the best fit as determined by quality of fit. The “correct” number of components is selected where the curve changes slope. The balanced scaling suggests 5 or 7 factors; the optimal scaling suggests 6 factors. (Reprinted from, Chemom. Intell. Lab. Syst., 18, Paatero, P. and Tapper, U., Analysis of different modes of factor analysis as least squares fit problems, 183-194.,1993, with permission from Elsevier.)

situation (Paatero and Tapper, 1994).

The second case involves the potential for “outliers,” or large values, to influence factors too strongly. These values can be individually scaled to reduce their impact on the factorization by adjusting their standard deviations (Paatero and Tapper, 1994). The statistical effect of such adjustments is not specified in the paper, and it may be wise to try several adjustments and compare the outputs from the method.
The final case involves missing data values, which can occur because of sample loss, laboratory error, failure of one piece of equipment in a set that is monitoring simultaneously, or other error. In this method, the missing value and its standard deviation can be replaced by the data and standard deviation column means, respectively. This ensures that the value will be roughly the same as the others in the column and it will not unduly influence the factorization. Though Paatero and Tapper (1994) state that in other factorization methods, missing data would require deletion of an entire row or column of the data matrix, it seems that this approach (or a similar one) could be used in other methods.

A final feature of PMF is an iterative reweighting which allows a robust analysis more tolerant of outliers. Because this is not a fundamental part of the iterative least squares solution it is not discussed further in this review, but it is an important quality of the PMF algorithm which is not shared by other source apportionment methods.

2.2 Positive Matrix Factorization PMF

2.2.1 The model

As described above, PMF seeks to fit a “source” and “factor” matrix to a set of data, following the model

$$E_{ij} = X_{ij} - Y_{ij} = X_{ij} - \sum_{h=1}^{p} G_{ih} F_{hj}$$  \hspace{1cm} (A4)

where $E_{ij}$ is an element of the residual matrix,

$X_{ij}$ is an element of the data matrix,

$Y_{ij}$ is the modeled element,

$G_{ih}$ and $F_{hj}$ are the score (source contribution) and factor (source profile), and
\[ i = 1, \ldots, m, \text{ and } j = 1, \ldots, n. \]

The object function to be minimized is then

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

(\text{A5})

where \( \sigma_{ij} \) is the standard deviation of the data element \( E_{ij} \).

The elements of \( G \) and \( F \) are constrained to be positive, reflecting the positive contributions (\( G_{ij} \)) to the linear combination of sources, which are comprised of positive concentrations of compounds (\( F_{ij} \)).

2.2.2 Extending the model to three dimensions

It is possible to expand the matrix factorization to three dimensions, such that Equation A4 becomes

\[ E_{ijk} = X_{ijk} - \sum_{h=1}^{o} A_{ih} B_{jh} C_{kh} \]  

(A6)

where \( i = 1, \ldots, m, \ j=1, \ldots, n, \) and \( k=1, \ldots, o \) and the object function \( Q \) in Equation A.5 becomes

\[ Q(E) = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{o} \left( \frac{E_{ijk}}{\sigma_{ijk}} \right)^2. \]  

(A7)

The three-dimensional case is neither the focus of this writing nor of the paper being reviewed, but further development of this model is described by Paatero in (Paatero, 1997b).
2.2.3 What versus how in the algorithm of PMF

Paatero makes the bold statement about the need to understand the inner workings of PMF:

“It is essential to stress that the technical details of the algorithm should in fact not concern the user of PMF. In the case of PMF, it is possible to separate the question ‘What is computed? i.e. which mathematical problem is solved by the program?’ from the question of how it is computed.”

While it may be true that understanding the algorithm is not necessary, such understanding would certainly enhance the ability of the user to interpret and defend the program’s results! Paatero’s point is not that the remainder of the paper can be discarded, but rather that this separation between method and result is not true of other methods. In PMF, standard linear algebra techniques give a least-squares fit to the data, and the link to the physical model resides in the non-negative quality of the factor and source matrices and the use of the measurement standard deviation in the minimized object.

2.2.4 The iterative PMF algorithm

The PMF model faces several challenges: non-linearity is created from the non-negativity constraints; residuals depend on products of unknowns; and the number of unknowns may be tens of thousands, so negatives cannot be removed one-at-a-time; the problem is usually ill-posed, so it may not have a unique solution. In order to tackle these obstacles, an algorithm was devised which utilizes alternating least squares methods.

In broad terms, the algorithm starts from pseudo-random initial values, in which the non-linearities are severe, so simple calculations are used for the first steps, changing one variable at a time. During the main steps, portions of the $G$ and $F$ matrices are held
constant while the others are solved iteratively. Between each iteration step, “rotations” are considered and outliers are scaled in robust mode. The calculations made during each iteration are described below.

### 2.2.5 Computation of the main steps

The goal of the main steps of the alternating least squares algorithm is to allow arbitrary matrices in the factor space of $G$ and $F$, $g$ and $f$, respectively, to change simultaneously while minimizing $Q(G+g, F+f)$. As indicated by the matrix addition, $g$ and $f$ must be sizes $m \times p$ and $p \times n$, respectively. The system then has $(m \times p + p \times n)$ or $(m+n)p$ unknowns; the convergence of the solution can be very slow. While this is not the approach used by the PMF algorithm, it is a straightforward description of a least-squares solution. This method, described by Paatero as a “full step” in the iteration because the full set of unknowns is allowed to change, takes the form

$$X = (G + g)(F + f) \quad \text{(A8)}$$

or $$X = GF + Gf + gF + gf. \quad \text{(A9)}$$

As $X-GF$ is final the error of fit ($E$), the error of fit at each step of the iteration could be termed the current residual, $R$. Then the iteration step is described by

$$R = Gf + gF + gf \quad \text{(A10)}$$

Paatero chooses to simplify Equation A10 by dropping the second order $gf$ term and using Gauss-Newton steps during the iteration.
It is possible to convert Equation A10 into a standard linear algebra form $Ax=b$, or in Paatero’s notation,

$$\Gamma \varphi = \text{vect}(R) \quad \text{(A11)}$$

where $\Gamma$ has dimensions $(mn \times (m+n)p)$ and contains many repetitions of values of the matrices $G$ and $F$,

$\varphi$ has dimensions $(m+n)p$ and contains all of the unknowns in $f$ and $g$, and

$\text{vect}(R)$ has dimensions $(m+n)p$ and contains all of the values of the residual matrix $R$.

If we try to imagine this system, recalling that $\Gamma \varphi$ reproduces $Gf + gF$, it might look something like

$$\begin{bmatrix}
G_{\text{row1}} & F_{\text{row1}}^T & \vdots & \vdots & F_{\text{rowp}}^T \\
G_{\text{row2}} & G_{\text{row2}} & \vdots & \vdots & F_{\text{rowp}}^T \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
G_{\text{rowm}} & G_{\text{rowm}} & \vdots & \vdots & G_{\text{rowm}} \\
\end{bmatrix}
\begin{bmatrix}
f_{11} \\
f_{21} \\
\vdots \\
g_{11} \\
\vdots \\
g_{1p} \\
g_{21} \\
\vdots \\
g_{mp} \\
\end{bmatrix}
= 
\begin{bmatrix}
G_{\text{row1}f_{\text{column1}}} + g_{\text{row1}F_{\text{column1}}} \\
G_{\text{row1}f_{\text{column2}}} + g_{\text{row1}F_{\text{column2}}} \\
\vdots \\
G_{\text{row1}f_{\text{columnp}}} + g_{\text{row1}F_{\text{columnp}}} \\
G_{\text{row2}f_{\text{column1}}} + g_{\text{row2}F_{\text{column1}}} \\
G_{\text{row2}f_{\text{column2}}} + g_{\text{row2}F_{\text{column2}}} \\
\vdots \\
G_{\text{row2}f_{\text{columnp}}} + g_{\text{row2}F_{\text{columnp}}} \\
\vdots \\
G_{\text{rowm}f_{\text{column1}}} + g_{\text{rowm}F_{\text{column1}}} \\
G_{\text{rowm}f_{\text{column2}}} + g_{\text{rowm}F_{\text{column2}}} \\
\vdots \\
G_{\text{rowm}f_{\text{columnp}}} + g_{\text{rowm}F_{\text{columnp}}} \\
\end{bmatrix}
$$

where empty spaces in the $\Gamma$ matrix represent zero row vectors of appropriate lengths. As Paatero notes, the $\Gamma$ matrix contains considerable duplication of the elements of $G$ and $F$ and numerous zeros. Because of its size and the duplication of values, these matrices are not useful for direct calculations, but they do form a useful conceptual model for a least-
squares fit. Equation A11 is the system which requires the least squares fit. The least squares solution is found by the following equations, with Paatero’s notation in the left column and Olver’s (2006) notation (from our text) on the right:

\[
\begin{align*}
\Gamma \varphi &= \text{vect}(R) \quad & A\mathbf{x} = \mathbf{b} & \quad \text{(A11)} \\
\Psi &= \Gamma^T\Psi' \quad & K = A^TCA & \quad \text{(A12)} \\
\gamma &= \Gamma^T W \text{vect}(R) \quad & \mathbf{f} = A^T C\mathbf{b} & \quad \text{(A13)} \\
\Psi \varphi &= \gamma \quad & Kx^* = \mathbf{f} & \quad \text{(A14)}
\end{align*}
\]

where \( C \) and \( W \) are positive-definite matrices of weights for the linear system. Paatero notes that Equation A14 may be solved by Cholesky decomposition into \( U^T U \) because \( \Psi \) is symmetric and positive definite; the \( U^T \) and \( U \) are, of course, invertible because they are of full rank, so the solution does exist.

Note that in the previous discussion of the “full step” solution, all of the values of \( g \) and \( f \) are unknown and allowed to change simultaneously! Handling so many variables is inefficient, so an alternate, “restricted step” is used for calculations. In this method, the number of variables allowed to change is “restricted.” Paatero describes this as follows:

First, the solution space is spanned by all components on the \( G \) side (\( mp \) unknowns), plus a single fixed direction on the \( F \) side. Altogether there are \( mp+1 \) unknowns to be determined. Second, the solution space is spanned by all components on the \( F \) side (\( np \) unknowns), plus a single fixed direction on the \( G \) side, altogether \( np+1 \) unknowns. (Paatero, 1997a)

The meaning of this is clarified somewhat by showing that in the restricted step, the simplified Equation A10 becomes

\[ R = \alpha Gf_{x-1} + g_x F \quad \text{(A15)} \]

where \( \alpha \) is a scalar which provides the “single fixed direction” of freedom on the \( F \) side by scaling its increment from the previous iteration step, \( f_{x-1} \),
\( G \) and \( F \) are fixed at their values from the previous iteration, and
\( g_x \) is the unknown matrix of increments to \( G \) in this iteration.

The unknowns are the matrix \( g \) (\( m \times p \) unknowns) and \( \alpha \), totaling \( mp+1 \). The next iteration would have residual

\[
R = Gf_x + \alpha g_{x-1}F
\]  

(A16)

where all of the matrix \( f \) is allowed to vary and \( \alpha \) acts as a scalar on the \( g \) matrix to increment \( G \). The matrix system in Equation A15 might look like

where empty spaces in the \( \Gamma \) matrix represent zero row vectors of appropriate lengths.

This strategy has two important benefits. First, this \( \Gamma \varphi = \text{vect}(R) \) system is significantly condensed from the full step version – in each iteration (Equation A15 or A16), the product \( Gf \) is always known; these values could be computed once and \( \Gamma \) is reduced to a size \( (mn)(mp+1) \) matrix which is essentially a block diagonal matrix, with
one full column in this orientation. In this smaller, better organized matrix, calculation of the Cholesky decomposition is simple and fast, improving the speed of the algorithm. The associated second benefit is the reduced computational memory required for holding the large matrices.

### 2.2.6 Computation of the rotational substeps

Paatero rightfully pointed out that what he calls rotations “…are not rotations in the proper mathematical sense. However, in the jargon of FA or PCA, all linear transformations are called ‘rotations’ and we are following this unfortunate practice” (Paatero and Tapper, 1993). A matrix $T$ and its inverse, $T^{-1}$, can be applied to the source and factor matrices $G$ and $F$ such that the residual of the fit, $E$, is not changed:

$$X = G T T^{-1} F + E = \tilde{G} \tilde{F} + E.$$

These “rotations” are applied between the iteration steps in such a way that an enhanced object function is minimized. (Despite its importance, this enhanced object function, which allows “automatic,” and therefore unbiased, rotations is beyond the scope of this review and is not discussed further. In essence, four additional terms are included in the minimization equation: two control the strength of the correction which keep $G$ and $F$ positive, while two others control the scalings on $G$ and $F$. Understanding the specific effects of these modifications to the minimization may be critical to justifying the rotations of solutions.)
2.3 The analysis of a synthetic test case with and without outliers

To demonstrate the interpretation of PMF results, Paatero provided a sample dataset which illustrated the concepts of solution rotation and handling outliers. Because this discussion of Paatero’s work did not include review of the robust outliers algorithm, only the solution to the non-outlier sample case is discussed here.

While PMF has been designed for use in exploring environmental data and an environmental example would interest his audience, Paatero cleverly points out that this would not make a good choice of example because the “correct result” is not known. This is his primary reason for creating a fairly simple dataset for exploring PMF’s capabilities. A second, and equally important, reason is that it is easiest to discern patterns from smooth curves than from less clean functions that might be found in a true environmental example. Paatero therefore creates an example that could represent a spectroscopic dataset of summed Gaussian peaks which decay exponentially over time to demonstrate his method.

2.3.1 The Gauss-exponential test case

Paatero created a set of four sources whose spectra are Gaussian distributions, each centered at a different wavelength; these make up the $G$ matrix. Each source decays exponentially at a different rate; these factors make up the $F$ matrix. The elements of $X$ are given by

$$X_{ij} = \sum_{h=1}^{4} a_h e^{-0.5(i-u_h)^2/w_h} e^{-\beta_h} + r_{ij} \quad \text{(A18)}$$

where $h$ is the number of the source,

$a_h$ dictates the magnitude of the contribution of the source,
the first exponential term gives the Gaussian peak shape,
the second exponential term gives the exponential decay of the source,
\( r_{ij} \) is a pseudorandom error with standard deviation = 0.01, intended to add noise to the data, and
the values of \( a, u, w, \) and \( t \) are given in Table A.1.

Paatero gives an example of his dataset, shown in Figure A.2. Careful inspection reveals the influence of four Gaussian peaks which decay at different rates. The parameters were selected so that rotation of the solution would be possible in order to demonstrate this important characteristic of PMF results.

2.3.2 Results: no outliers in the matrix

For the purposes of this example, Paatero only examined PMF results for the correct number of factors, so no discussion of the choice of correct number of factors is

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Component Number (( h ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>( a )</td>
<td>100</td>
</tr>
<tr>
<td>( U )</td>
<td>10</td>
</tr>
<tr>
<td>( W )</td>
<td>2.5</td>
</tr>
<tr>
<td>( T )</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Figure A.2.** Selected “spectra” from the synthetic dataset. Careful inspection reveals the influence of four Gaussian peaks in the total spectra. The $t_i$ labels mark the selected spectra displayed from the complete timeseries. As time increases, the spectral intensity decreases, reflecting the exponential decay built into the model. (Reprinted from, Chemom. Intell. Lab. Syst., 37, Paatero, P., Least squares formulation of robust non-negative factor analysis, 23-35.,1997, with permission from Elsevier.)

included. The results of a non-rotated, four-source decomposition are shown in Figure .

The source profiles, on the left of the figure, and their respective contributions, on the right, do not look like the expected Gaussian peaks and exponential decays. Because the “correct answer” is known in this case, it is clear that this is not the desired result and a rotation should be attempted. The source profiles are sums of single Gaussian peaks, so a rotation which “subtracts from $G$, adds to $F$” would approach the desired solution.

When PMF is allowed to make automatic rotations (included in the previous discussion only briefly), the results in Figure A.3 are achieved. Now the matrices have the expected result, with fairly clean, Gaussian peaks on the source side and clear exponential decay on the contribution side.
**Figure A.3.** Non-rotated results of the Positive Matrix Factorization of the synthetic dataset. Curves on the left represent the source profiles of the four sources; these correspond to the rows of the $G$ matrix. Curves on the right represent the corresponding contributions of these four sources; these correspond to the columns of the $F$ matrix. Because we know the “correct” answer, it is clear that this is not the proper rotation of the solution to achieve the expected shapes (labeled above the figure) of the sources and factors. (Reprinted from, Chemom. Intell. Lab. Syst., 37, Paatero, P., Least squares formulation of robust non-negative factor analysis, 23-35.,1997, with permission from Elsevier.)

Paatero notes that “the ability to recover the true solution by this automatic rotation technique depends critically on the fact that each true $G$ factor contains many values which are close to zero or exactly zero.” This helps to push the solutions toward one edge of the rotational domain, hopefully helping to identify a “correct” rotation of data that is not so smooth.

---

3 Paatero’s original figures included values on the axes, but these values were purely arbitrary. In the figures displaying the results, the source and factor matrices were plotted on the same graph with a continuous scale in the x-axis, but this has no real connection to the two separate types of data. These arbitrary scales have been removed in this presentation and replaced with generic labels to clarify the meaning of the graphs. As expected, scales increase in the vertical direction for the y-axis or to the right for the x-axis.
**Figure A.3.** Rotated results of the Positive Matrix Factorization of the synthetic dataset. Curves have the same representations as those in Figure 1. Rotation was made such that subtractions were favored on the source side (the $G$ matrix, left) and additions favored on the contribution side (the $F$ matrix, right) to achieve the expected shapes (labeled above the figure). (Reprinted from, Chemom. Intell. Lab. Syst., 37, Paatero, P., Least squares formulation of robust non-negative factor analysis, 23-35, 1997, with permission from Elsevier.)

2.4 Availability of the Programs

Trial versions of the PMF2 and PMF3 (two- and three-dimensional versions, respectively) and the associated documentation are available from Paatero’s FTP site, ftp://rock.helsinki.fi/pub/misc/pmf/ and at ftp://ftp.clarkson.edu/pub/hopkepk/pmf, a US mirror of the main site.

3. Discussion and Conclusions

The Positive Matrix Factorization method solves a least-squares fit to a system of matrices, thereby extending the least squares fit of a vector discussed during this course. The method also incorporates linear transformation and rotation as discussed in this course. These rotations are constrained by the need to maintain positivity in all of the
values of the “source” and “factor” matrices which comprise the original data matrix. The use of measurement standard deviation to normalize the data matrix improves the connection between the mathematical and physical models and also appears to give an optimal weighting to the least squares system. Though it was not discussed in this review, this weighting is an important aspect of handling outliers in the data, a common occurrence in a “real” dataset.

This powerful application of linear algebra is a respected technique for exploring data matrices in a variety of environmental research areas, as is demonstrated by its extensive citation in the literature. Because the solutions may not be unique, careful defense of a particular rotation is required, but the choice of rotation can have a significant mathematical, and therefore unbiased, basis. Many researchers present results of multiple source apportionment methods together to explain their choice of result to strengthen this defense.
Appendix B

PMF Evaluation Tool (PMF)

B.1 Introduction

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

The ipfs were written by Ingrid Ulbrich and Donna Sueper (Jimenez Group, University of Colorado, Boulder) and Greg Brinkman (Hannigan Group, University of Colorado, Boulder). Questions about this code can be addressed to Ingrid or Donna. PMF (Positive Matrix Factorization) was developed by Dr. P. Paatero (Dept. of Physics, University of Helsinki).

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

B.2 Installing PMF with Igor

B.2.1 PMF and Operating Systems

The PMF executable is compiled only for Windows/DOS. The PET has generally been tested by Donna and Ingrid in Windows XP, with some testing in Windows Vista. It should run
well on either platform; please contact Donna if you suspect you have operating system problems with the PET.

For users with Macs, two options have proven successful. One method is to execute PMF on a Windows computer and then analyze the experiment on a Macintosh. The other method is to execute PMF under the Windows emulator on a Mac. Note that in this latter case, you may need to have the PMF executable somewhere in the C:\ directory; this isn't necessary when running Windows on a PC.

B.2.2 Setting up PMF on Your Computer the First Time

It is recommended that you use the BareBones version to make sure that PMF and Igor interact correctly on your computer.

1. Download the BareBones Starter Kit
2. Create a new folder on your computer where you'll store the PMF files and all files output by PMF (this should NOT be the folder where you store your Igor experiments).

   This folder must contain:
   • PMF2wtst.exe (obtained from P. Paatero or the BareBones Starter Kit)
   • imupmf.ini (obtained from the BareBones Starter Kit)
   • pmf2key.key (obtained from P. Pattero, U. Helsinki)

3. Start a new Igor experiment and load the following files from the BareBones Starter Kit:
   • DataAndErrorForBareBonesPMF.itx
   • BareBonesPMFExecution_1_00C.ipf

B.2.2.1 Running the PMF Test Case (BareBones Version)
1. Start the Panel from the BareBonesPMF menu on the menu bar.

2. Use the first button and select the path to the folder you created with the PMF2wtst.exe file.

3. Use the second button to execute PMF.

You should see a black DOS window pop up, scroll a lot out output, and then close. Igor will then tell you, "The PMF barebones analysis was successfully completed within X seconds."

\[ X \text{ should be } > 0! \]

If the window goes away immediately and the Igor message says that the analysis was completed within 0 seconds, the execution was not successful. Follow the steps for what to do If PMF does not Run Properly to correct the problem.

**B.2.3 Setting Up the Full Version**

After you have confirmed that PMF and Igor interact correctly on your computer with the BareBones version, you are ready to use the full version for analysis.

1. Download the .ipf (or .pxt) files from the software section of the wiki.

2. Create a new folder on your computer where you'll store the PMF files and all files output by PMF (this should NOT be the folder where you store your Igor experiments).

This folder must contain:

- a version of the PMF executable file (obtained from P. Paatero or the BareBones Starter Kit, see also an important note about PMF2.exe files)
- mypmft.ini (obtained from the software section of this wiki)
- pmf2key.key (obtained from P. Pattero, U. Helsinki)
3. Start a new Igor experiment and load the 3 .ipf files or load the template. Create your data and error matrices, load them to your experiment, and then proceed to the PMF analysis.

4. If there are problems executing PMF, follow the steps for what to do If PMF does not Run Properly to correct the problem.

B.2.3.1 Important Note about PMF2___.exe Files

PMF executable files are available from P. Paatero in versions that have been compiled for different operating systems. Executable file names that include $w$ are compiled for Windows.

At the present time (Nov. 1 2009) all versions of PMF2___.exe are believed to work with the PMF Evaluation Tool. It is important to use the file mypmft.ini for v2.03 or later with files such as pmf2wopt.exe. These newer .exe files check the maximum line length defined in the .ini file, which the old .exe files did not seem to do.

In the headers of the PMF_Execution_v2_03.ipf (and later versions) file, check these lines for the .exe file you have:

```
strconstant PMF_EXEC_FILE_NAME="pmf2wtst.exe"

//strconstant PMF_EXEC_FILE_NAME="pmf2wopt.exe" //imu2.03
NOTE that if you use this line, you need to upgrade your .ini file to v2.03!!
```

B.2.3.2 If PMF does not Run Properly

1. Look in the folder you created with the PMF2wtst.exe file for the existence of the files Matrix.dat and StdDev.dat
If these files exist, Igor was able to access the correct folder. Continue with Step 2.

If these files do not exist, Igor was not able to access the correct folder. Go back to the first button on the panel and check that you've given the correct path to the folder with the PMF2wtst.exe file. Run PMF again by pressing the second button.

2. Look in the folder you created with the PMF2wtst.exe file for the existence of the file PMF2.LOG

If this file does not exist, PMF was not run in this folder. Go to step 3.

If this file does exist, PMF attempted to run in this folder. Open the file PMF2.LOG (it is a text file).

Glance down the contents of the file and look for many lines of sequential numbered output, such as

1 rank1 step chi2=   9282.6  Penalty= 1.5287E+04 Flags GF
2 rank1 step chi2=   7411.7  Penalty= 1.4084E+04 Flags GF

If these lines are in the file, PMF ran successfully on your computer, which must be fast enough to run this data in less than 1 second. You're done! Everything works and you can proceed with real data.

If the lines of sequential numerical output are not in the file, you should find the following lines within this file (note that every line is NOT included here, but the lines selected here are in the order they appear in the file):
2a)

```text
###PMF2 .ini file for: IMUPMF.INI --- BareBonesPMF

Successfully read task initialization file imupmf.ini titled:  ###PMF2 .ini file for: IMUPMF.INI --- BareBonesPMF

If these lines appear in the file, PMF found the .ini file. Continue with Step 2b.

If these lines do not appear, you should see a message about not finding an appropriate .ini file.

Make sure that this folder contains the file imupmf.ini, provided with the BareBones PMF Starter Kit. If it did not, copy the file to this folder, delete the file PMF2.LOG and press the second button in the panel again to see whether PMF runs successfully.

If the file imupmf.ini already exists in the same folder as the PMF2wtst.exe file, continue with Step 3.
```

2b)

```text
Successfully opened input file 30 with name MATRIX.DAT

Successfully opened input file 31 with name STD_DEV.DAT
```
If these lines appear in the file, everything should have run correctly. Look at the rest of the PMF2.LOG file to see whether other errors are reported. If you still encounter difficulty, contact Ingrid for assistance and attach the PMF2.LOG file to your email.

If these lines do not appear in the file, you will see a message about PMF not being able to access one of these files. Check that the files are not being used by other programs, delete the file PMF2.LOG, and press the second button on the panel again to see whether PMF runs successfully.

3. Look in the C:\ directory of your computer (C:Documents and Settings\..\My Documents with the full version) for the existence of the file runpmf.bat

If this file exists, Igor was able to write to your C:\ drive. Continue with Step 3a.

If this file does not appear, Igor was not able to write this file to your C:\ drive. This might be due to high security settings on your computer. You should create a text file with this name (NOT runpmf.bat.txt) and choose to edit it (not open it) with a text editor (e.g., Notepad, WordPad, Emacs, etc.). The file should contain the lines

```bash
cd C:\Documents and Settings\Ingrid Ulbrich\Desktop\ pmf2wtst imupmf
```

NOTE that you must change the path in this example to the path to the folder where you have put the file PMF2wtst.exe!

3a)
Execute `runpmf.bat` by double-clicking on its icon. You should see the black DOS window pop up, scroll output, and close again.

If this happens, PMF has run successfully from the batch. In the Igor experiment, press the second button on the panel to see whether PMF runs successfully.

If the black window does not show scrolled output, continue with Step 3b.

3b)

You now need to execute PMF from the command line.

To access the DOS command line, launch the Command Prompt from your Windows Start menu, located in

```
Start -> Programs -> Accessories -> Command Prompt
```

Change from the path displayed in the prompt to the folder where you put the `PMF2wtst.exe` file. Use the command

```
cd
```

to change directories (e.g.,

```
cd Desktop\PMF
```

). You can change one directory at a time, or several at a time as shown in the example.

To move up 1 directory, use

```
cd ..
```

Be sure that the folder contains the three files
by listing the contents of the current directory, using the command
dir

Then type this command at the prompt to run PMF:

pmf2wtst imupmf

You should see scrolling output in the command window. The window will not disappear and you can scroll back through the output (which is also saved in the file PMF2.LOG).

If the output contains many lines of sequential numbered output, such as

1 rank1 step chi2= 9282.6 Penalty= 1.5287E+04 Flags GF
2 rank1 step chi2= 7411.7 Penalty= 1.4084E+04 Flags GF

PMF ran successfully here. Delete the file PMF2.LOG and in the Igor experiment, press the second button on the panel to whether PMF runs successfully from Igor.

If the output does not contain these lines, go back to Step 2 of this section to examine errors that might be reported by PMF in the file PMF2.LOG.

B.3 Creating the Organics and Error Matrices (Step 0)

B.3.1 Creating the Matrices, by Instrument (Software) type
The Igor AMS PMF tool needs 4 inputs: two 2-dimensional matrices of the same size and two 1-dimensional waves corresponding to indexes of the rows and columns. One 2-dimensional matrix is the organics matrix (typically PMF is only performed on organics), the other is the error of the organics matrix. The Igor PMF code also needs two 1-dimensional waves corresponding to index values for the rows and columns of the data matrix. These are the times series (corresponding to the rows) wave and a m/z wave (corresponding the columns).

Regardless as to whether the data has been generated from a quadrupole or tof ams, user should select the use MS airbeam correction option for generating the organic matricies. This is because the error matricies will be generated using the airbeam correction factor if it exists.

Some steps are slightly different between manipulation of the quadrupole and the tof ams data sets. In the quadrupole the removal of spikes and the smoothing of data may be necessary, where as in the tof data sets, neither may be required (due to the fact that the quadrupole only samples one m/z at a time).

B.3.1.1 In the Q-AMS Software (James')

Be sure to use v1.41 or later of the Q-AMS Analysis Software ("James' Program"). Corrections have been made since earlier versions to the error calculation routines!

Download from Qi Zhang's website Extract Waves&matrices v 1.1.ipf (note that v1.2 is for Squirrel/HR data) and include it in your experiment with James' software.

Call org_mats, which will calculate a data matrix (organics_MS) and error matrix (organics_MS_err) in root: in your expemeint, and save these matrices along with the timeseries for organics, sulfate, nitrate, ammonium, and chloride in a file called
"WavesMatricesForOrganicAnalysis.itx" (Igor should prompt you for the folder where you want to save the data). All of the data are saved in ug/m3 with all corrections (CE, RIE) applied.

You may want to load the saved waves into a new experiment to run PMF.

You'll also need to include your time series wave (t_series) and a wave of the m/z's in the matrix (amus).

*Continue to the Deleting NaNs/zeros for all Instruments, section B.3.2.3*

**B.3.1.1.1 Recommended Practice: Removing Spikes**

"Spikes" in the time series of an m/z can occur in Q-AMS data from large but infrequent particles during the scanning of the quadrupole. If such spikes have a common source with a factor that can be retrieved by PMF, they may increase the variation of that factor profile and additional factors may be found that represent this variation, but not a physically-meaningful, separate component. The "excess signal" from these spikes can be subtracted from the spikes and the average mass spectrum of the spikes examined. See Zhang et al (2005).

Note that if you remove "excess signal" in the method of Zhang et al. (2005) and leave the error values for these points unchanged, you are automatically "downweighting" these points in PMF. This is appropriate because the replacement value for the original spike is not known as well as the values for points without spikes.

**B.3.1.1.2 Optional Practice: Smoothing**

Smoothing can be used to reduce high-frequency noise in the data that could also be fit as additional factors. If you smooth the data, be sure to propagate this smoothing in the error matrix (not added to the wiki yet).
3.1.2 In SQUIRREL

In Squirrel the main two data sets containing the unit resolution sticks and unit resolution errors bare MSSDiff_p and MSSDiff_p_err. These data sets may reside in the root folder or may reside in intermediate files, depending on the version of squirrel and the settings chosen by the user. The code within squirrel will automatically retrieve these data sets as needed.

If the error matrix has not yet been calculated, this must be done first. In the Corrections tab, Errors sub-tab, check the calc MS errors checkbox. Make sure you don't intend to do any other calculations in the corrections tab and then press the Do Corrections button. This will generate the MSSDiff_p_err data set.

The Squirrel interface for generating the Org and Org error matrices in Squirrel have changed slightly before and after version 1.46 - 1.48. In all versions one goes to the MS tab, average mass spectra section, enters (only) Org as a species. In versions 1.46 and higher one needs to check the 'Calc, plot 1sigma err' checkbox and optionally use the 'Prompt for max m/z' checkbox and then press 'Calc Time Series Spectra' button (v. 1.47+). In versions 1.46 and lower one uses the 'Export Matrices' button.

After the operation is finished the matrices you need will reside in the root folder: MSSD_xxx_Mat_Org and MSSD_xxx_Mat_Org_err (v. 1.46 and lower) or MatD_xxx_Org and MatD_xxx_Org_err (v. 1.47) where xxx is the name of the todo wave. Both of these matrices will have zeros in the m/z columns where no organics contribution is identified (as is given in the frag_organics wave) and blanks in the rows for run numbers not included in the todo wave. In
the case of version 1.46 and lower, the MSSD_xxx_Mat_Org_err will not be calculated unless MSSDiff_p_err has first been created.

Now you have the two most important basic pieces - the Org matrix and the Org error matrix.

The two 1-dimensional waves corresponding to the rows and columns of the matrix are the time series wave and a simple wave corresponding to m/z. In squirrel the time series wave is root:index:t_series. There is a m/z wave, root:diagnostics:amus whose values are simply column numbers, 1,2,3,...1000. This is the wave that is plotted as the x axis when one generates and average mass spectrum. For simplicity, the maximum value is always 1000, regardless as to what the maximum m/z data value is. Users can either redimension this amus wave to be the correct size (the same number of points as the column of the data matrix) or to create a new wave simply by the command line.

There is still more work that needs to be done: calculating the minimum error, removing columns with zeros/nans, and perhaps doing some editing on the error matrix.

*Continue to the Deleting NaNs/zeros for all Instruments, section B.3.2.3*

**B.3.1.3 In PIKA**

Pika has evolved rapidly between 2008-2010, and generating the High Resolution stick matrix in Pika will depend on the version you are using.

**3.1.3.1 Pika 1.08 and higher**

The interface and behavior for generating in memory HR sticks and errors is very similar to that in Squirrel.
As in the case for Squirrel, the Igor AMS PMF tool needs 4 inputs: two 2-dimensional matrices of the same size and two 1-dimensional waves corresponding to indexes of the rows and columns. However in Pika one will typically want two 1-dimensional waves describing the HR ions within the columns: one indicating the mass value (i.e. 12.000) and one indicating the chemical formula (i.e. C etc).

There is a button in the HR Results tab in the lower left corner titled "Calc time series spectra (2d matrix, not plotted, leave list blank for all HR ions)". Most AMS users will want to analyze the HR Org species, so include this item in the control for selecting the results (right under "Step 6."). From the drop down menus, most will want to choose the OMinusC data type, check the Calc, plot err. checkbox, and select the OMinusClIonCountErr error type. When the user presses the "Calc time series..." button a 2-D data matrix called called HRMxM_HROrg (or HRMx*) and a 2-D error matrix called HRMxM_HROrg_err (or HRMx*err) will be generated.

HR data and error matrices will have the number of rows equal the the root:index:t_series wave, just like in squirrel. Rows will have nan values for runs not in the todo wave selected. If the HROrg species was selected (typical), all calculations in the HR batch table and HR_frag_organic have been applied to both matrices, just as in squirrel.

Whereas in squirrel matrix columns correspond to integer m/z values (column 0 = m/z 1 etc), for high resolution spectra, columns correspond to individual HR ions. So if the first three HR ion fit were C, j13C, CH, then the first column is C, the second is j13C, the third is CH, and so on. If an HROrg matrix was generated there may be column for HR ions that were not fit but were included in the HR frag table, such a S j33S, etc. The text and numeric waves of HR ions corresponding to the columns are root:HR_frag:SpeciesMassText and
root:HR_frag:SpeciesMassWave. For users of 1.90+ a button called "HR ions for HR species table" will bring up these waves in a table for you. You will often want to export two waves to indicates HR column values - one containing the exact mass values (12.00000, etc) and one containing the HR ion chemical name (C, etc).

In Pika 1.09F there is a handy checkbox that when checked will automatically generate an itx containing all the waves necessary for the PMF (with the exception of any data preparation steps such as the minimum error, Nan removal, etc.)

B.3.1.4 For non-AMS Users

You will need to create 5 waves to run your PMF analysis in Igor:

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

This section describes how you might make those waves, assuming your data is in Excel and is loaded to Igor with one wave for each species and another wave for each species' error values. Depending on the initial format of your data, some steps may not apply to you.

B.3.1.4.1 Loading data from Excel into Igor

In Igor, go to the Data pulldown menu and choose Load Waves -> Load Excel File... (Note that the Excel file must be closed when you try to load the waves.) You can choose the
wave names based on column headers from the Excel file if you wish. Note that you can probably uncheck "Make double precision waves."

If your Excel file has data values in one sheet and error values in another (especially if the columns have the same titles), you may wish to load all of the data waves in one datafolder and all of the error waves to another datafolder and form the data and error matrices in these locations.

Note that in Igor you cannot use a digit as the first character in a wave name, nor may you use operators (+ - * / ? etc.) or spaces in wave names.

B.3.1.4.2 Concatenating 1-D Waves to form a 2-D matrix

You will use Igor's concatenate function to make the 1d waves into a 2d matrix. This is easiest to do with a "wavelist" -- a string that contains all of the relevant wave names, separated by a semicolon (the default separator in Igor). For example,

```
string/G SpeciesWaveNameList = wavelist("", ";", ")
```

will make a string called SpeciesWaveNameList with a list of all of the waves in the current folder. You may wish to check that all of the wave names are correct in the string. You may need to remove any wave names from the list that you don't want to include in the matrix (e.g., year). This can be done by a line like

```
SpeciesWaveNamesList = removeFromString("year;month;day;PM10;", 
SpeciesWaveNamesList)
```

Now you can make the matrix by
concatenate SpeciesWaveNameList, DataMx

When you do this for the error matrix, be sure to check first that the strings in the data and error folders are identical so that all species are included in both waves in the same order!!

**B.3.1.4.3 Making a Text Wave of Species Names**

The SpeciesWaveNamesList can be converted to a text wave using a function in the Igor PMF software (so these .ipf's must be loaded in the experiment:

```plaintext
gen_list2txtWv(SpeciesWaveNamesList , "SpeciesWaveNames")
```

The wave of species names is now called SpeciesWaveNames. (You'll only need to do this once, since the list is the same for the data and error waves.)

**B.3.1.4.4 Making a Numeric Wave of Species Indexes**

You can create a numeric wave with indexes counting from 0 by

```plaintext
make/N=(numpnts(SpeciesWaveNames)) SpeciesWaveIndex = p
```

(Note that "p" is an Igor convention for indexes. You can count from 1 by changing "p" to "p+1".)

**B.3.1.4.5 Making a Time Series Wave**

Igor has its own convention for time, counting in seconds from 1/1/1904. Because this creates very large numbers, time waves should be double precision (/D flag). Two examples for making time waves are shown here.

For daily data, with input waves year, month, and day:
make/N=(numpnts(year))/D timeseries = date2secs(year,month,day)

For data with input waves year, month, day and also hour, minute, and second:

make/N=(numpnts(year))/D timeseries = date2secs(year,month,day) + hour*60*60 + minute*60 + second

B.3.2 Further Preparation of the Error Matrix

The following steps are recommended for AMS datasets and follow the practices laid out in Ulbrich et al., ACP, 2009 (with more detailed references in each section below). Note that the only mandatory step is Deleting NaNs/zeros. The functions for the error modifications can be found in pmf_ErrPrep_AMS_v2_3.ipf. More extensive documentation on use of the functions is given in the headers in that file.

Before running these functions, you should duplicate your error matrix and give it a short name (fewer than 12 characters). Each function lengthens the wavename and the functions will complain if the name gets too long.

B.3.2.1 Recommended Practice: Set a Minimum Error

Ions arrive at the mass spectrometer detector with a Poisson distribution. The error for a counted number of ions is sqrt(counted number of ions). The smallest number of ions we can count in one run is, of course, zero ions, but perhaps there was one and it was missed. The error for counting zero is sqrt(0), but an error of 1 would be more appropriate in this case. Hence a minimum error threshold of 1 ion is set.

The minimum error is applied in three steps:
1. In the experiment with James' panel or Squirrel, calculate the signal equal to 1 ion with the function pmf_err_minErr1ion_ugm3 in pmf_ErrPrep_Q-AMS_OneIonEquiv_v2.3.ipf or pmf_ErrPrep_ToF-AMS_OneIonEquiv_v2.3.ipf. This function operates on the current todo wave.

2. Copy the wave minErr1ion_ugm3 or minErr1ion_Hz (depending on the units of your matrix) into your PMF experiment.

3. Run the function 'pmf_err_errMx_minErr (found in pmf_ErrPrep_AMS_v2_3.ipf), using your error matrix with the short name.

The function produces the following waves:

1. The error matrix with minimum error applied called nameOfWave(errMx)+"_min" (e.g., OrgMSerr becomes OrgMSerr_min)

2. A matrix of the fractional increase of the errors called nameOfWave(errMx)+"_adjErrMask" where the value of each point is (new/old)-1.

See also discussion of Ulbrich et al., ACPD 2008, P. Paatero comment (p. S5730) and Author response (p. S11960)

**B.3.2.2 Propagation of Smoothing (when relevant)**

Any smoothing of the data matrix must be propagated in the error matrix. The function pmf_err_propogateSmooth propagates box or Gaussian smoothing.

Some notes about specifying the smoothing that was performed for the data:

1. Allowed types of smoothing are "box" and "Gaussian".

2. The type of smoothing that is done in the AMS software is selected in the Misc tab.
3. The number of points used in box and Gaussian smoothing is used as defined in Igor.
   - For box smoothing, the number of points refers to the size of the box. I.e., smoothing that includes 1 adjacent point on each size is 3-point smoothing.
   - For Gaussian smoothing, the number of points refers to the number of adjacent points used. I.e., smoothing that includes 1 adjacent point on each side is 1-point smoothing.

The function produces a wave with the propagated error called nameOfWave(errMx)+"Prop" (e.g., OrgMSerr_Min becomes OrgMSerr_minProp).

**B.3.2.3 Deleting NaNs/zeros for all Instruments**

The matrices produced in the previous steps have NaNs in all rows from bad runs and 0 values in columns with good runs that have no organic fragments. All of these rows and columns need to be removed before running PMF. This is a two-step process.

1. First, change the columns with 0's to NaNs.
   a. You can do this by changing all 0's to NaNs, e.g.

   \[
   \text{OrganicMx} = \text{OrganicMx}[p][q] == 0 ? \text{NaN} : \text{OrganicMx}[p][q]
   \]

   b. You could make an organic fragments mask wave (= 1 for organic fragments, = NaN for others) and multiply your matrix by the wave. This is also a good time (and easy way) to delete other columns you may not want to retain in the PMF analysis (e.g., m/z's 19 and 20, which are small copies of m/z 44 in the normal frag table).
Note that 1b is safer because of the way this function works (see *Important Note* below)! If an actual good data value in the first row was 0 (unlikely but possible) and you have replaced zeros by NaNs, a column may be deleted inadvertently!

2. In the PMF_Execution_XX.ipf, use the function

\[
\text{pmf\_ams\_deleteNaNs\_mxWvs(dataMx, errMx, rowDescrWv, colDescrWv)}
\]

where

\[
\text{dataMx and errMx are your data and error matrices,}
\]

\[
\text{rowDescrWv is a 1-D wave that gives the indices for the rows (usually t\_series), and}
\]

\[
\text{colDescrWv is a 1-D wave that gives the indices for the columns (usually amus).}
\]

The function creates new versions of the input waves with names noNaNs\_amus, noNaNs\_t\_series, etc. from which the NaN rows and/or columns have been deleted. 2-D waves called NaNsWv\_amus and NaNsWv\_tseries have been created and are used by related functions to delete items from other waves or reinsert the original NaNs into waves. (Note: NaNsWvs were stored as NaNsLists before v2.03.)

3. Check whether any NaNs remain in the matrix. (This may happen with HR data where the fit returned NaN for a particular peak.) You must replace any NaNs with some value and choose an accompanying error value. It is imperative that you describe how these values were handled and how many values were effected when publishing results!

(People who have addressed this issue are encouraged to add to the recommendations in this section!)
Considerations for Replacing NaN Values

- These values are suspected to be equal to or close to 0. Zero values are allowed in the PMF input, but cannot actually be fit as zero. Very small positive values are probably a better choice than zero.
- Read and evaluate other studies performed with PMF. This issue is more common in filter datasets than in AMS data and several practices have been developed for dealing with below detection limit and missing values in these dataset. Consider whether the practices in the literature are appropriate for AMS data.

Considerations for Setting Uncertainty Values for Replaced NaNs

- The uncertainty of the replaced NaN value should be estimated from the uncertainty of that fragment.
- The estimated uncertainty for the replaced NaN value should be increased by a factor of ~100 so that this point has almost no weight.
- As above, read and evaluate other studies performed with PMF and consider whether practices in the literature are appropriate for AMS data.

B.3.2.3.1 Important Note about Using DeleteNaNs Functions!!

The pmf_ams_deleteNaNs_mxWvs function chooses the first row and column of the data that contain a mix of NaNs and values and uses this as a template to delete rows or columns (respectively) that contain NaNs. It does not check every single value in the matrix. This is because there should not be any nans or infs in the matrix unless there are nans or infs in the entire row or column. After running this function, it may be wise to check (e.g., with wavestats) whether all NaNs have been removed.
Here is a piece of code that can remove columns with zeros. This approach presumes you have already used the pmf_ams_deleteNaNs_mxWvs function above. (Copy and paste these lines into a procedure window and execute from the command line.)

// A function for removing columns in a matrix that have only zeros

// For example, if a user has an "Org" matrix, it would remove columns
// corresponding to m/z 1 - 11, 14, etc.

// sample usage: RemoveZeroCols(noNaNs_mx, noNaNs_mxerr, noNaNs_amus, NaNsList_amus) // where noNaNs_mx and noNaNs_mxerr are replaced with your wave names

Function RemoveZeroCols(noNaNsMx, noNaNsMxerr, NoNansAmuWave, NoNansAmuList)

wave noNaNsMx, noNaNsMxerr, NoNansAmuWave

string NoNansAmuList

variable idex, numRow, numCol

numRow = dimsize(noNaNsMx,0)

numCol = dimsize(noNaNsMx,1)
make/o/n=(numRow) tempCol

for (idex = numCol; idex>=0; idex-=1) // work 'backwards' so that we don't mess up the counting
    tempCol = noNaNsMx[p][idex]

wavestats/q/m=1 tempCol

if (V_min ==0 && V_max == 0)

    deletepoints/m=1 idex, 1, noNaNsMx, noNaNsMxerr

    deletepoints/m=1 idex, 1, NoNansAmuWave

    NoNansAmuList = num2str(idex+1) + ";"+ NoNansAmuList // keep adding at the front to keep in numerical order.

endif

endfor

killwaves/z tempCol

End

B.3.2.3.2 Related Functions

Also found in PMF_Execution_XX.ipf:

- pmf_ams_deleteNaNs_Wvs(wvList, NaNsList)
Applies the NaNsList_amus or NaNsList_t_series to delete points from any list of waves of the same dimension.

- \texttt{pmf\_ams\_deleteNaNs\_Mxs(wvList, NaNsList, NaNsDimension)}

  Applies the NaNsList_amus or NaNsList_t_series to delete points from any list of matrices in the specified dimension (using standard Igor dimensions, where 0=rows, 1=columns). This could be used e.g. to delete the same set of rows or columns from a matrix other than the data matrix (used as the original template) or the error matrix. Not currently available in \texttt{PMF\_Execution\_XX.ipf}; contact Ingrid if you're interested in this function.

- \texttt{pmf\_ams\_insertNaNs\_Wvs(wvList, NaNsList)}

  Applies the NaNsList_amus or NaNsList_t_series to insert points into any list of waves of the same dimension. This is very helpful for making final t_series waves for publication that don't have lines across periods of no data.

- \texttt{pmf\_ams\_insertNaNs\_Mxs(wvList, NaNsList, NaNsDimension)}

  Applies the NaNsList_amus or NaNsList_t_series to insert points into any list of matrices in the specified dimension (using standard Igor dimensions, where 0=rows, 1=columns).

- \texttt{pmf\_ams\_NaNsList\_2\_NaNsWv(NaNsListName)}

  Converts NaNsLists (v. 2.02 and earlier) to NaNsWvs (v 2.03).

- \texttt{pmf\_ams\_RemoveBlockSize\_NaNsWv(NaNsWv, BlockSizeToRemove)}
Creates NaNsWv_type_trunc and deletes the records for original blocks of NaNs of size BlockSizeToRemove. If you did V-W switching and are only running the V-mode data in PMF, your NaNsWv_tseries will have an entry for every other row of your original matrix noting that 1 NaN row was removed. You may also have periods with longer contiguous runs of NaNs (e.g., instrument was down, filter runs, etc.). For publication plots, you should insert NaNs into your final factor waves so that periods of missing data are not crossed by a line connecting the good data before and after the real gaps. For NaNs removed due to menu switching, however, you may not want to insert all of these NaNs again. (In the example case, each solved value would have a NaN on each size of it and you couldn't plot with Lines Between Points. If you uncheck "Gaps" to connect these points, you'll also have the undesirable lines across periods of missing data.) Running pmf_ams_removeBlockSize_NaNsWv(NaNsWv_tseries, 1) would create NaNsList_tseries_trunc which would not include a record of all the removed W runs and then could be used to insert NaNs representing only missing data periods in the final waves.

B.3.2.4 Recommended Practice: Downweight "Weak" Variables (m/z's)

Any m/z's that have low signal-to-noise ratio (SNR) may, in fact, have more noise than signal. If these m/z's contribute enough Q, PMF tries to fit the noisy data. In this way, the inclusion of such m/z's can be detrimental to the PMF analysis. If the error associated with these m/z's is increased, the Q-contribution (residual/error) is decreased, "downweighting" these points' contribution to the fit. m/z's with SNR<0.2 are considered "bad" by Paatero and Hopke
(2003) and should be removed or strongly downweighted (factor of ~10). m/z 's with 0.2<SNR<2 are considered "weak" and should be downweighted (factor of 2-3).

The calculation of SNR and downweighting of "weak" m/z's is carried out in three steps:

1. Calculate SNR of each m/z using function pmf_err_SNRwv using the data matrix, the version of the error matrix generated in the previous step, and the model error that will be used in the panel. The function generates a wave of the SNR for each m/z called nameofwave(DataMx) + ",_SNRwv".
2. Check the graph produced for "bad" m/z's. These are not removed in the next function. To remove these columns, you'll need to rerun the DeleteNaNs step after making them into NaNs or changing them in the mask wave. (This is better than just deleting the columns by had because they'll be added to the NaNsList_amus and will be reinserted if you insertNaNs later.
3. Downweight "weak" m/z 's with function pmf_err_DwntWeakColumns using the SNRwv generated in the previous step, the error matrix used to calculate the SNRwv, and the multiplicative value used to downweight the weak m/z 's (Paatero and Hopke recommend 2-3).

The function generates a new error matrix called nameofWave(errMx)+"Wk" (e.g., noNaNs_orgMSerr_minProp would become noNaNs_orgMSerr_minPropWk).

See also Paatero, P., and Hopke, P. K (2003).

B.3.2.5 Recommended Practice: Downweight Peaks Related to m/z 44 in Frag Table
In the default fragmentation table, the information in m/z 44 is repeated 6 or 7 times (in m/z's 16, 17, 18, 19, 20, (28 in the Aiken et al. 2008 revision), and 44) with different proportionalities. PMF fits correlations, regardless of the magnitudes of the signals. Repeating the information of m/z 44 several times implies that it's really (x6 or 7) important, which it isn't! It is possible to downweight the columns of these m/z's so that in total they only contribute the m/z 44 signal once. (It would be possible to remove the repeated information and replace those columns after running PMF, but we think that downweighting them is logistically simpler.)

Downweighting the m/z's related to m/z 44 is accomplished in two steps:

1. Make a wave that contains the m/z's related to m/z 44 (in any order), e.g.

   make/N=4 mz44peaksWv = {16, 17, 18, 44}

   make/N=5 mz44peaksWv = {16, 17, 18, 28, 44}

   make/N=6 mz44peaksWv = {16, 17, 18, 19, 20, 44}

   make/N=7 mz44peaksWv = {16, 17, 18, 19, 20, 28, 44}

2. Use function pmf_err_dnwt44peaks with the error matrix generated in the previous step, the mz44peaksWv, and the noNaNs_amus wave.

   The function generates a new error matrix called nameofWave(errMx)+"44" (e.g., noNaNs_orgMSerr_minPropWk would become noNaNs_orgMSerr_minPropWk44).

See also Supp. Info. for Ulbrich et al. (2009)

B.4 Perform PMF Analysis *Step 1*

B.4.1 Additional File for PMF Executable Directory
Be sure that the file `mymft.ini` (provided on the PMF data page) is located in the directory with the PMF Executable.

**B.4.2 Recommended Practice: Organizing Your Folders**

`+ root`

`+ TemplateData` to copy for new versions of analysis

`+ Variation1` e.g., your basis case

`+ Variation2` e.g., with different error estimates

`+ External_MassSpectra` for use with the Scatter Panel

`+ External_Tseries` for use with the Scatter Panel

NOTE: Data for running PMF can be in root: or a subfolder of root:, but not any lower folder.

**B.4.3 The Executable Panel**

1. Set the path to the PMF executable
2. Provide the data and error matrices information
   a. Chose the folder (must be root: or a directory in root:)
   b. Choose the Data and Error matrices (use noNaNs_ versions)
   c. Choose model error (PMF increases the errors provided by `newError = oldError + modelError*dataValue`)
3. Choose the type of PMF analysis
a. Exploration will run PMF for a range of number of factors and FPEAKs or SEEDs. This is the typical use for exploring a dataset and comparing many solutions.

b. Bootstrapping explores the uncertainty of one solution (i.e., one number of factors at one fpeak for one seed). This is usually a final step run only on the solution you have selected from the exploratory analysis.

4. Choose a range for number of factors.
   a. When checking to make sure that everything runs properly, you may want to run just one case (Min p = 2, Max p = 2).
   b. Recommended Practice: Run cases with 1 factor to have a context for the meaning of the 2-factor solution.
   c. In the Bootstrapping mode, only the "min p" is read; the "max p" is ignored.

5. Choose FPEAK or SEED values
   a. "FPEAK" is a tool used to explore rotations of the solutions of a given number of factors. Note that FPEAK does not explore all possible rotations of a solution. FPEAK = 0 does not apply any rotational forcing. Non-zero values of FPEAK create near-zero values in the factor profiles (mass spectra) or time series. More information about FPEAK can be found in the PMF Users Manual Part 1 (pp. 9,12,14,21) and Part 2 (p. 24), and in several papers by P. Paatero (see Other Resources).
      i. A good first set of FPEAK values is -1.0 to +1.0 with a delta value of 0.1 or 0.2. For a full analysis, a wide enough range of FPEAKs to achieve Q/Qexp of at least 1% above the minimum value is recommended.
ii. In Exploration mode when varying the fpeak, the Seed is set in headers in the PMF_Execution...ipf file: constant DEFAULT_SEED = 0

iii. In Bootstrapping mode, only the "min fpeak" is read; the "max fpeak" is ignored.

b. "SEED" is a tool used to choose different random starts (initial values) for the PMF algorithm. Using different seeds may lead to solutions in different local minima (Q/Qexp) in the solution space. One set of solutions may have more physical meaning than another, or multiple sets may make physical sense. It is impossible to test all start values, but testing many seeds may give an indication of local minima for your dataset. More information about seeds can be found in the PMF Users Manual Part 1 (p. 11) and Part 2 (p. 16).

i. Run seeds from 0 to your preferred maximum with a delta value of 1.

ii. In Exploration mode when varying the Seed, the fpeak is set in the headers of the PMF_Execution...ipf file: constant DEFAULT_FPEAK = 0

1. Exploring Seeds with a non-zero fpeak should be done with caution, as the "push" of a magnitude of fpeak for one area of local minima may be different than the "push" of the same magnitude of fpeak for a different area of local minima.

iii. In Bootstrapping mode, the Seed is set in headers in the PMF_Execution...ipf file: constant DEFAULT_SEED = 0

6. Select checkboxes

a. Run PMF in background Each execution of PMF (see Exploration Mode Summary) creates a black DOS window that pops up. If the box is not checked,
this window "grabs the focus" and makes itself the top window. This makes it
hard to use the computer for anything else. If the box is checked, the window will
not grab focus, but Igor and your computer's CPU will be busy.

B.4.3.1 What the Software Does When You Press the Button...

B.4.3.1.1 Exploration Mode Summary

The software will execute PMF once for every combination of number of factors and
FPEAK/seed. So if you run 1-5 factors and 5 FPEAKs, PMF will run 5x5=25 times. Each run
starts a new black DOS window that will close when the run is completed. The duration of each
run is printed in the history at the end of each run. In general, runs which solve for more factors
and runs with FPEAK farther from 0 take longer. The code runs all of the FPEAKS or seeds for
one number of factors, then advances to the next number of factors (e.g., run 1 factor with each
of 5 FPEAK values, then 2 factors with each of 5 FPEAK values, etc.).

A little more detail

The software writes the files

\[ \text{C:\delete\_log.bat} \quad \text{C:\runPMF.bat} \]

and writes your DataMatrix and ErrorMatrix as MATRIX.DAT and STD\_DEV.DAT,
respectively to the folder with your PMF Executable. The software also writes a file to that
folder called STD\_DEV\_PROP.DAT, which has the same number of points as the DataMatrix
and in which every element is equal to the ModelError.

The software then enters a pair of nested loops in which the following steps occur:
• for each number of factors
  • for each FPEAK or SEED
    - use the file `mymft.ini` as a template to create the file `imupmf.ini`, which is used as the control file for PMF.
    - In v2.03, the convergence criteria for completing the PMF calculations can be set proportional to Q/Qexp by changing headers in the `PMF_Execution...ipf`. The default is to use the convergence criteria in `mymft.ini`.

    //Settings for PMF Iteration Convergence
    proportional to Qexp

    constant ITERATION_CHI2_PROP_QEXP = 0
    constant FIRST_ITER_LEVEL_PROP_CONST = 2e-6
    constant SECOND_ITER_LEVEL_PROP_CONST = 2e-6
    constant THIRD_ITER_LEVEL_PROP_CONST = 1e-6

    You can read more about convergence criteria and large datasets in the PMF Users Manual Part 1 (pg. 10) (see Other Resources).

  o Delete the old PMF2.LOG file by running `delete_log.bat`
  o Execute PMF by running `run_PMF.bat`.
  o Wait for PMF to complete its run.
  o Load PMF output (including log file and factors)
At the completion of the loops, the software calculates some statistics from the output and then creates a panel to select data for viewing.

4.3.1.2 Bootstrapping (added in v2.02)

The bootstrapping mode is developed after the method described in the EPA PMF v3.0 Users Manual Sect. 6.4 (see Other Resources). The bootstrapping method is used to estimate the uncertainty in both the factor mass spectra and time series. This is achieved by running PMF on the full dataset once (from one PMF solution (i.e., combination of number of factors and fpeak) at a time) and then making a series of variations (the number is specified by the user when selecting the bootstrapping mode) in which a subset of the original rows (mass spectra) are randomly replaced by other rows from the original matrix and running PMF on each of these.

For each new PMF case ("bootstrapped case"), the resultant factors are compared to those from the original dataset and assigned as "matching" the original factor with which it has the highest correlation. Bootstrapped cases in which each bootstrapped factor was matched to exactly one of the original factors (i.e., there is a one-to-one mapping between original factors and those from the individual boot-strapped cases) are retained for calculation of the average mass spectrum and time series of bootstrapped factors. Plots of the original factors and the average bootstrap factors with 1-sigma variation bars are produced automatically.

EPA PMF Users Manual recommends doing 100 bootstrap runs for final results.

A little more detail

All output from the bootstrapping runs is saved in folder root:pmf_bootstrap:. 
Row (mass spectrum) replacement is performed by using the StatsResample function in Igor to select rows for replacement. The row values are then sorted in increasing order as a convenience.

- The 2d wave RowsToBeReplaced records the rows to be used in each bootstrapped case. Each column represents one bootstrap case. Each column lists the rows of the original matrix included in that bootstrapped case.
- The 2d wave ReplacementHistogram counts the number of times that each original matrix row was used in a bootstrap case. Each column represents one bootstrap case. Summing the rows of this matrix gives the total number of times that each original matrix row was used in the bootstrapping cases.

The assignment of bootstrapped factors to the factors from the original case is made by Pearson R correlation. Factors are assigned only on the basis of mass spectral comparison, and each factor is assigned to one of the original factors.

- Note that this is different than in EPA PMF. Our code does not have a criterion for the lowest allowable correlation between bootstrapped and original factors. In EPA PMF, factors that fall below this limit are "unmapped"; no factors are "unmapped" in our code.
- Note that if the original case has factors that are very similar to each other, the assignment of the bootstrapped factors may be incorrect or ambiguous. No current work has been done to give guidance as to what "very similar" means. No sanity checks are made in the code for this type of situation.
• The 3d wave FactorProfile_Rval stores the correlation between the boostrapped and original factors. Rows represent the factors from the original case, columns represent the factors from the bootstrapped cases, and layers represent each bootstrapped case.

• The 2d wave FactorProfileSort contains the number of the original factor to which each boostrapped factor (row) has been matched in the bootstrapped case (column). Columns which contain (e.g., in a case with three factors) "0, 1, 2" have factors which were uniquely matched to the columns in the original case; these will be included in the averages of mapped factors. Columns which contain duplicate entries (e.g., "0, 1, 0") have multiple factors that were matched to the same original factor; these cases will not be included in the averages of the mapped factors.

Note that the criteria for including bootstrapped cases in calculations of the average bootstrapped factors is different in our code than in EPA PMF. Bootstrapped cases in which two or more factors are mapped to the same original factor are not included in the averages in our code. In these instances, the whole bootstrapping case is rejected before calculating averages. In EPA PMF, all bootstrapped factors mapped to the same original factor are included in the average.

**Important:** At the present time (v. 2.02) the plots produced by the bootstrapping code expect to find the waves noNaNs_amus (for plotting profiles) and noNaNs_t_series (for plotting time series). If these are not the names of your waves that describe the columns and rows, respectively) of the input matrix, you should duplicate your row and column description waves to have these names.

**B.4.3.2 Running Two Simultaneous Analyses on Dual-Processor Computers**
You can run two PMF analyses (in two separate experiments) on the same computer simultaneously if you have dual processors. Each analysis will run at the same speed as on a single-processor computer (or when one analysis is run on the dual processor computer). The PMF executable is not "multi-processor aware," meaning that it can not utilize both processors simultaneously for one PMF run.

To run two simultaneous analyses with the PET, you'll need two directories on your computer with the PMF .exe, .key, and mypfmt.ini files. The directory names must end in "1" and "2," respectively. For example, you could have

C:\PMF\PMF Executable1

    pmf2wtst.exe

    pmf2key.key

    mypfmt.ini

C:\PMF\PMF Executable2

    pmf2wtst.exe

    pmf2key.key

    mypfmt.ini

Each experiment running PMF must use a separate Executable directory. While a PMF analysis associated with a PMF Executable directory is running, a file called "PMFrunning.txt" exists in that directory. If you try to run your second PMF analysis using the same Executable
directory, the PET will give you an error. You can choose a different Executable directory and press the button to start the analysis.

B.5 View PMF Analysis Results *Step 2*

The View Results panel has controls to change which solution is being displayed (upper left corner) and shows plots (left-to-right and top-to-bottom) of

- Factor fractional variance and mass contribution
- RotMat
- Q vs. number of factors in the solution
- Q vs. FPEAK
- Factor time series
- Factor profile
- Time series, reconstruction, and residual for the current species
- Total time series and reconstruction
- Box-and-whiskers plot of scaled residuals for each species
- Scaled residual histogram for the current species
- Total residuals, total absolute residuals, total fractional residuals, total fractional absolute residuals, and total Q contribution as a function of time or species
- Correlation of factor profiles and time series in the current solution

B.6 Compare PMF Results with External Factors *Step 3*

Caution: This part of the code is a bit fussy! Please contact Ingrid if you have tried the tips here and have trouble getting this panel to work.
B.6.1 Setting up the External Data

1. Check that the waves NaNsWv_amus" and "NaNsWv_tseries" (these were created when you deleted NaNs from your original matrix) are in the datafolder where your PMF output is being saved. If they're not already there, you'll need to find them and copy them to this location.
   a. Versions 2.02 and lower used the strings "NaNsList_amus" and "NaNsList_t_series" to record the location of deleted NaNs. If you are upgrading experiments, make sure that these strings are in the folder where your PMF output is being saved; "NaNsWv_amus" and "NaNsWv_tseries" will be created and used for future calculations.

2. You'll need separate folders (in root: ; they cannot be in a lower directory) for mass spectra and time series you want to use for comparison to the factors.

3. The tricky part of using this panel is setting up your mass spectra and time series correctly.
   a. Each wave must have either the same number of points as in the corresponding dimension of your noNaNs_data matrix or the same number of points as in the corresponding dimension of your original matrix.
      i. For example, if your original matrix is 3246 rows and 300 columns and your noNaNs_matrix is 3200 rows and 268 columns, your "external" mass spectra can have 300 or 268 points; "external" time series can have 3246 or 3200 points.
   b. For mass spectral comparisons, download "full" spectra (usually 300 points) from the AMS Spectral Database instead of using the shortened ones provided in the
9th Users Meeting template. You should inspect the length of all waves from the Database to make sure that every one has the correct number of points for your work.

4. IMPORTANT NOTES

a. The reason for the restriction on the number of points in "external" comparison waves is the following: After you select the datafolders for the external mass spectra and time series, the code makes a folder inside each of these folders called "noNaNs". Each wave in the external datafolder is copied to the new directory. Then the code checks whether the waves have the same number of points as the same dimension in the matrix used to run PMF; if so, no change is made to the wave. If not, the code _assumes_ that the wave has the dimension of the original matrix (which it doesn't know about) and therefore uses the string "NaNsList_amus" or "NaNsList_t_series" to delete the rows that it believes were NaN in the original dataset. (It's ok if these waves still have NaNs (e.g., missing points in data from another instrument when the AMS data was good); only the points where both the factor and external waves have valid data are included in the correlation calculation.)

b. Because of some internal coding restrictions, time series waves for comparison cannot include the string "series" in their name; such waves will not be created in the noNaNs folder.

c. Each of the mass spectral and time series waves must be 1-dimensional. This means that you cannot use the waves from root:pmf_plot_globals: called TseriesFactor1, TseriesFactor2, etc. since these are 2-dimensional waves.
d. Old versions of the mass spectra in the AMS Spectral Database have m/z = point number, meaning that point 0 = m/z 0. This is not usually the way that AMS matrices are saved from James' software or from Squirrel, so you may need to delete 1 point from the beginning of each spectrum that you use from the Database. This is important to check, or else you correlate the wrong m/z's with each other.

B.6.2 Choosing the Folders with External Data

1. The first time you want to calculate factors you can do so by pressing the "External Data Panel" button on the main panel or by choosing from the PMF pulldown menu "Compare PMF results with External Factors *Step 3*".
   a. Note that after you have accessed the "External Data Locations" panel, pressing the "External Data Panel" button on the main panel will not bring you back to the selection panel again. To choose different folders or force recalculations you must access the selection panel from the PMF pulldown menu.

2. Select your external data folders.
   a. Other choices from the pulldown menus include
      i. "No external data of this type": The PET will not attempt to calculate correlations between factors and external data of this type.
      ii. "Update List": If after calculating factor correlations you wish to add new external waves of this type, choose this option to add the correlations for the new waves do your existing list of correlations.

3. Press the button to proceed.
B.6.3 What the PET Does (and how to fix things if something goes wrong)

1. In each external data folder, a folder called "noNaNs" is created as described in the IMPORTANT Note 1 above (item 3 of Setting Up the External Data).

2. Each of these new waves is compared to every factor wave. (This can take a while if you have a lot of waves for comparison.)

   a. If the factor and external waves have different lengths, the function aborts and tells you that the comparison function was called with waves of different lengths. Unfortunately, it doesn't give helpful information about which wave had the wrong length (we'll try to look into that and improve that error message).

      i. If this happens, you should look in the "noNaNs" folder in the appropriate external data folder and check whether all of the waves have the correct length. Waves with incorrect numbers of points in this folder may be the result of incorrect wave lengths in the "external data" folder. Try to fix all of the problem waves and then run the function for calculating the scatter comparison again by choosing step 3 from the PMF pull-down menu. Recall that this is the only way to force recalculation of the comparison of the factors!

         1. If forcing recalculation didn't fix the problem, delete the "noNaNs" folder in the appropriate external data folder and then recalculate again.

3. The correlation values between the factor waves and the external data waves are stored in waves in the folder with PMF output called

   i. "RcorrMx4d_Profiles" and "RcorrMx4d_Tseries" (with Pearson R)
ii. "RcorrMx4d_Profiles_pear_mzGrt44" (with Pearson R, only for m/z > 44)
   
b. It is also possible use the function scat_calc_RCorrMx4d_UC() to calculate
   
i. "RcorrMx4d_Profiles_UC" and "RcorrMx4d_Tseries_UC" (with the
   Uncentered Correlation, as reported in Ulbrich et al. (2009)

4. When the calculation is complete, the Scatter Panel is created.

B.6.4 Other Potential Problems and Solutions

*Pulldown menus don't have lists.* Each "noNaNs" folder should also contain a text wave called "TseriesWvsNms" for time series or "FactorWvsNms" for profiles. This wave is used for the pulldown menus in the panel. If this wave is missing, the pulldown menus may not work. There should also be a string of wave names in the folder called "TseriesWvsNmsList" or "FactorWvsNmsList"; if so, you can create the text wave by gen_list2txtWv(listStr, wvNm).

B.6.5 Some Other Notes

*Order of the factors.* Factors are numbered 1 to N and match the factors in the main panel, counting from the bottom of the factor plots.

*Colors.* Factor 1 is black, Factor 2 is red, Factor 3 is green, Factor 4 is blue, etc. Factors in this panel have the same color as they did in the main panel. In the overlay plots, the factor is its usual color and the external species is orange.

*Size of Factor Space and Current Fpeak value sliders.* The sliders in this panel and in the main panel control both panels simultaneously. Graph updates are slower with both panels. Be patient and don't click on anything until everything has updated.

B.6.6 More Features
Assign Groups to External Data. This feature allows you to reorder the external data waves and assign them to groups. Groups then define the colors used in the R bar plots in the panel and in the Comprehensive External Data Correlation Plot (below).

Comprehensive External Data Correlation Plot. This plot display the "R vs External Factor" plots for all factors at once.

B.7 Considerations for Choosing a Solution

Recommendations about choosing solutions are outlined in Chapters II and III.

B.8 PMF Evaluation Tool Software

At the time of this writing, the most recent version of the PET is 2.03A, which can be found at http://cires.colorado.edu/jimenez-group/wiki/index.php/PMF-AMS_Analysis_Guide.
Appendix C

Calculation of Error Values for Synthetic Datasets

The calculation of error values for real and synthetic data is described schematically in Fig. C.1. In real datasets, the error, $\sigma_{\text{diff}}$, for the total difference signal (open beam – closed beam) at each point is estimated by

$$\sigma_{\text{diff}} = \alpha \sqrt{\frac{I_o + I_c}{t_s}}$$

where $\alpha$ is a factor of 1.2 applied to account for the random variation of the areas of single-ion signals, $I_o$ and $I_c$ represent the signal, in ion Hz, of the open and closed beams, respectively, and $t_s$ is the time, in seconds, spent sampling each $m/z$ (Allan et al., 2003). The error for a particular species (e.g., organics), $\sigma_{\text{species}}$, is calculated by propagating the application of a “fragmentation matrix” (Allan et al., 2004) to the total difference error $\sigma_{\text{diff}}$, such that

$$\sigma_{\text{species},ij} = \sqrt{\sum_{k=1}^{n} \left( \sigma_{\text{diff},ik} x_{\text{Fragmentation Matrix}}_{kj} \right)^2}.$$  

(C2)

Note that in the synthetic case, neither the open beam nor closed beam values are known, so $\sigma_{\text{diff}}$ and $\sigma_{\text{species}}$ cannot be calculated directly. We therefore proceed through the following steps to estimate errors for synthetic data. The synthetic organics matrix, $X_{\text{input}}$, is converted from $\mu g/m^3$ to Hz using the campaign-averaged conversion factor from the real Pittsburgh dataset. The matrices of inorganic species and air (in Hz) are added to the organics matrix to give a total estimated difference matrix. The calculation of the $\sigma_{\text{diff}}$ matrix (Eq. C1) requires an estimated closed beam signal matrix, the estimated open beam signal, and the sample time per $m/z$. The closed signal is estimated as the campaign-averaged closed mass spectrum. The open signal is
Figure C.1. Schematic diagram for calculation of error matrices for real (black path) and synthetic (blue path) AMS data.

Estimated as the estimated total difference matrix (representing open – closed) plus the campaign-averaged closed mass spectrum. The sample time per m/z is estimated as the campaign-averaged sample time. Calculation of $\sigma_{\text{diff}}$ for the synthetic case uses $\alpha = 1.2$. The calculation of the $\sigma_{\text{organic}}$ matrix (Eq. C2) uses the organics fragmentation matrix from the Pittsburgh dataset. The $\sigma_{\text{organic}}$ matrix is then converted from Hz back to $\mu g/m^3$ using the inverse of the campaign-averaged conversion factor used above. Comparison of the magnitude of the errors of the real dataset and the synthetic dataset is shown in Fig. C.2, indicating that this
procedure results in noise estimates very similar to those from the real dataset. The synthetic errors are generally 20% lower than the real errors because of the difference in the applied $\alpha$ values.

![Figure C.2. Comparison of errors calculated for the Pittsburgh dataset (y-axis) and the synthetic two-factor dataset (x-axis). Markers are labeled by their $m/z$ and colored by time during the campaign (beginning is red, end is blue). The black line is a 1:1 line. Outlying red points are from the first sample (row of the matrix), made over less than the standard 5-minute averaging time, decreasing the error from these points when using the longer campaign-averaged sample time. The large band of green outlying points is mainly from $m/z$ 41, which had much higher than average background levels in the instrument during two events during the campaign.](image-url)
Appendix D
Supporting Information for Chapter II

Weighting Factors for m/z 44-Related Peaks

The organic mass spectrum measured by the Q-AMS is estimated from the full measured mass spectrum but application of a “fragmentation table” (Allan et al., 2004). Several of the m/z’s in the AMS organic spectrum are related to m/z 44 by proportional constants in the fragmentation table such that the signals, x, are related by

\[ x_{m/z \ 18} = x_{m/z \ 44} \]  \hspace{1cm} (D1)

\[ x_{m/z \ 17} = 0.27 \ x_{m/z \ 18} = 0.27 \ x_{m/z \ 44} \]  \hspace{1cm} (D2)

\[ x_{m/z \ 16} = 0.04 \ x_{m/z \ 18} = 0.04 \ x_{m/z \ 44} \]  \hspace{1cm} (D3)

and likewise their errors have the same relationships:

\[ \sigma_{m/z \ 18} = \sigma_{m/z \ 44} \]  \hspace{1cm} (D4)

\[ \sigma_{m/z \ 17} = 0.27 \ \sigma_{m/z \ 44} \]  \hspace{1cm} (D5)

\[ \sigma_{m/z \ 16} = 0.04 \ \sigma_{m/z \ 44} \]  \hspace{1cm} (D6)

Including all of these m/z’s gives duplicates the signal at m/z 44 and gives excessive weight to this large signal in the PMF fit. The group of duplicate peaks (m/z’s 16, 17, and 18) could be omitted from the PMF input and then reconstructed afterwards, but it would be more convenient to include them in the fit but weigh them so that the Q-contribution of the group is the same as the Q-contribution if only m/z 44 had been included.

In order to satisfy

\[ Q_{m/z \ 44 \ only} = Q_{all \ m/z \ 44-related \ ions} \]  \hspace{1cm} (D7)
assume that the errors when all m/z 44-related peaks are included must be multiplied by a factor $a$ such that

$$
\sum_{j=1}^{m} \left( \frac{X_{m/z \ 44}}{\sigma_{m/z \ 44}} \right)^2 = \sum_{j=1}^{m} \left( \frac{X_{m/z \ 44}}{a^* \sigma_{m/z \ 44}} \right)^2 + \sum_{j=1}^{m} \left( \frac{X_{m/z \ 18}}{a^* \sigma_{m/z \ 16}} \right)^2 + \sum_{j=1}^{m} \left( \frac{X_{m/z \ 17}}{a^* \sigma_{m/z \ 16}} \right)^2 + \sum_{j=1}^{m} \left( \frac{X_{m/z \ 16}}{a^* \sigma_{m/z \ 16}} \right)^2 \quad \text{(D8)}
$$

Substituting Eqs. (SI-1) to (SI-6),

$$
\sum_{j=1}^{m} \left( \frac{X_{m/z \ 44}}{\sigma_{m/z \ 44}} \right)^2 = \sum_{j=1}^{m} \left( \frac{X_{m/z \ 44}}{a^* \sigma_{m/z \ 44}} \right)^2 + \sum_{j=1}^{m} \left( \frac{X_{m/z \ 44}}{a^* \sigma_{m/z \ 44}} \right)^2 + \sum_{j=1}^{m} \left( \frac{0.27X_{m/z \ 44}}{a^* 0.27\sigma_{m/z \ 44}} \right)^2 + \sum_{j=1}^{m} \left( \frac{0.04X_{m/z \ 44}}{a^* 0.04\sigma_{m/z \ 44}} \right)^2 \quad \text{(D9)}
$$

all of the proportionality constants cancel, and the equation can be reduced to

$$
\sum_{j=1}^{m} \left( \frac{X_{m/z \ 44}}{\sigma_{m/z \ 44}} \right)^2 = \frac{4}{a^2} \sum_{j=1}^{m} \left( \frac{X_{m/z \ 44}}{\sigma_{m/z \ 44}} \right)^2 \quad \text{(D10)}
$$

and $a = 2$ (i.e., $a$ = the square root of the number of m/z 44-related ions).

Analyses of HR-ToF-AMS data (Aiken et al., 2008) have led to revisions of the fragmentation table such that m/z 28 is now assigned organic mass. When the revised fragmentation table is used, the scaling factor $a$ for the error should be sqrt (5) = 2.24 because there are five ions (m/z’s 44, 28, 18, 17, 16) related to m/z 44.
Calculation of the Angle between a Vector and a Plane

The factor MS can be considered as vectors in a 270-dimensional space. Any two vectors in this space define a plane. Here we consider the plane defined by the HOA and OOA-1 MS vectors. Any 3rd vector in that 270-dimensional space will have an angle with respect to that plane. Any vector which is the linear combination of HOA and OOA-1 would lie exactly in that plane and have an angle of zero with respect to the plane. Any vector not in the HOA-OOA-1 plane is not a linear combination of HOA and OOA-1, and its “distance” from the HOA-OOA-1 plane can be quantified by the angle $\alpha$ between the plane and the vector.

In linear algebra terms, let $A$ be a 270 x 2 matrix formed by combining the non-orthogonal, non-collinear column vectors $HOA$ and $OOA-1$ such that

$$A = (HOA, OOA-1)$$ (D11)

and rank$(A) = 2$. A projection matrix $P$ may be constructed by

$$P = A(A^T A)^{-1} A^T$$ (D12)

which gives the projection $(w)$ of a third vector $(v$, which is OOA-2 in the quoted example) onto the plane by

$$P v = w$$ (D13)

A right triangle between $v$ and the HOA-OOA-1 plane can be identified in which $w$ is the base (and lies in the HOA-OOA-1 plane), $v$ is the hypotenuse, and a vector $z$ could be drawn perpendicular to the plane to define the height (Olver and Shakiban, 2006). The cosine of the angle $\alpha$ between $V$ and the plane can be defined by

$$\cos \alpha = \frac{\|w\|}{\|v\|}$$ (D14)

and $\alpha$ can be calculated with the inverse cosine function.
Figure D.1. Comparisons of the uncentered correlation with Pearson R. All PMF factors are from solutions of the real Pittsburgh case with solutions from 1 to 7 factors at FPEAKS from -3.0 to +3.0. Sixty MS from the Mass Spectral Database are used for MS comparisons; 34 tracer TS are used for TS comparisons. a) MS from PMF solutions vs. database MS; b) MS from PMF solutions vs. MS from PMF solutions; c) TS from PMF solutions vs. tracer TS; d) TS from PMF solutions vs. TS from PMF solutions.
Figure D.2. Solutions of the real Pittsburgh dataset in robust mode (thicker black markers) and the non-robust mode (thinner colored lines and markers, in front) with a) 2 factors, b) 3 factors, and c) 4 factors.
Figure D.3. PMF solutions of the real Pittsburgh case with a) 2 factors, b) 4 factors, and c) 5 factors.
Figure D.4. $R^\text{MS}$ between representative spectra from the AMS Mass Spectral Database (http://cires.colorado.edu/jimenez-group/AMSSd) and a) the third factor mass spectrum from the 3-factor PMF solution of the real Pittsburgh dataset, b) the fourth factor mass spectrum from the 4-factor PMF solution of the real Pittsburgh dataset, and c) the “mixed” factor mass spectrum from the 3-factor PMF solution of 2-factor base case. Values are given in Table D.1. Superscripts denote the source of the reference spectra as follows: (a) Zhang et al., 2005; b) Canagaratna et al., 2004; c) Alfarra et al., 2004; d) Alfarra, 2004; e) Bahreini et al., 2005; f) Sage et al., 2007; g) I.M. Ulbrich, J. Kroll, J.A. Huffman, T. Onash, A. Trimborn, J.L. Jimenez, unpublished spectra, FLAME-I, Missoula, MT, 2006; h) Schneider et al., 2006)
Figure D.5. Diurnal profiles of a) OOA-2, Nitrate, and Chloride, and b) $\frac{Q}{Q_{\text{exp}}}$ and the total residual from the 3-factor solution of the real Pittsburgh dataset.
Figure D.6. Solutions of the real Pittsburgh dataset from 64 random starts computing a) 3 factors, b) 4 factors, and c) 5 factors.
Figure D.7. Results from bootstrapping analysis of the 3-factor solution of the real Pittsburgh case. Average (black) with 1-σ error bars (grey) are shown for factor a) MS and b) TS. The solutions from multiple FPEAKS (Fig. 10) show a greater range in MS than the 1-σ variation bars, while the TS show a similar range to the 1-σ variation bars.
Figure D.8. a) OOA-1, OOA-2, and HOA time series from the 3-factor solution of the real Pittsburgh dataset. b) Total Residual, c) Total absolute residual, and d) absolute residual normalized by total signal for the 3- to 6-factor solutions of the real Pittsburgh case.
Figure D.9. Histograms (red) of scaled residuals of selected m/z’s from the 3-factor solution of the real Pittsburgh case. The first bin includes all scaled residuals < -10; the last bin includes all scaled residuals > 10. A Gaussian distribution (mean = 0, $\sigma = 1$, multiplied by 3199 to have the same area as the histogram) is shown in black to guide the eye.
Figure D.9 cont.
Figure D.9 cont.
Figure D.10. Change in $Q/Q_{\text{exp}}$ in the real Pittsburgh dataset from 3 to 6 factors vs. a) OOA-1 contribution, b) OOA-2 contribution, and c) HOA contribution from the 3-factor solution.
Figure D.11. Correlations of PMF factors to each other as they change with FPEAK for a) the real Pittsburgh case and b) the three-factor synthetic base case. The red labels denote the correlations of the input factors. Traces are colored by FPEAK and numbers denote the FPEAK of each solution. Black dots in b) indicate the correlation of the factors in the input. c) Correlation of the PMF factors to the input factors for the three-factor synthetic case. d) Expansion of c).
Figure D.12. PMF solutions of the 2-factor synthetic base case with a) 2 factors, b) 3 factors, c) 4 factors, and d) 5 factors.
Figure D.13. PMF solutions of the 3-factor synthetic base case with a) 2 factors, b) 3 factors, c) 4 factors, and d) 5 factors.
Figure D.14. 3-factor solutions of the 3-factor synthetic base case for selected “good” FPEAK values.
Figure D.15. a) Correlations between 3-factor real output and all synthetic input cases. b) Retrieval correlations between PMF and input TS and MS versus mass fraction of different third factors in synthetic cases. Markers denote the resemblance of the factors to the input MS or TS. For the MS, O, S, F, and B refer to OOA-2, SOA, FA, and BBOA, respectively. For the TS, T refers to the input TS. For both cases, H and I refer to HOA and OOA-1, respectively.
Figure D.16. Correlation versus FPEAK between PMF factor a) MS and selected reference MS, and b) TS and tracer TS.
Table D.1. Correlations between PMF factor TS and tracer species.

<table>
<thead>
<tr>
<th></th>
<th>$U_{TS}$</th>
<th>$R_{TS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOA-1 vs. Sulfate</td>
<td>0.95</td>
<td>0.85</td>
</tr>
<tr>
<td>OOA-2 vs. Nitrate</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>OOA-2 vs. Chloride</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>HOA vs. CO</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>HOA vs. NOx</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>HOA vs. EC</td>
<td>0.93</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Table D.2. Correlations between PMF factor and selected reference MS from the AMS spectral database.

<table>
<thead>
<tr>
<th>Reference Spectrum</th>
<th>a) OOA-2, Real Data</th>
<th>b) OOA-1a, Real Data</th>
<th>c) &quot;mixed&quot; Factor, Synthetic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3-factor solution</td>
<td>4-factor solution</td>
<td>3-factor solution</td>
</tr>
<tr>
<td></td>
<td>UC, RMS m/z &gt; 44</td>
<td>UC, RMS m/z &gt; 44</td>
<td>UC, RMS m/z &gt; 44</td>
</tr>
<tr>
<td>HOA Pittsburgh (Zhang et al.)</td>
<td>0.74 0.91 0.71 0.91</td>
<td>0.37 0.69 0.32 0.65</td>
<td>0.82 0.98 0.92 0.99</td>
</tr>
<tr>
<td>Diesel Bus Exhaust</td>
<td>0.90 0.93 0.89 0.93</td>
<td>0.68 0.63 0.65 0.59</td>
<td>0.92 0.97 0.96 0.98</td>
</tr>
<tr>
<td>Lubricating Oil</td>
<td>0.74 0.92 0.71 0.93</td>
<td>0.32 0.54 0.27 0.50</td>
<td>0.76 0.93 0.86 0.95</td>
</tr>
<tr>
<td>OOA Pittsburgh (Zhang et al.)</td>
<td>0.83 0.56 0.82 0.47</td>
<td>0.96 0.94 0.96 0.93</td>
<td>0.86 0.71 0.71 0.59</td>
</tr>
<tr>
<td>aged rural</td>
<td>0.84 0.78 0.83 0.73</td>
<td>0.98 0.96 0.98 0.95</td>
<td>0.87 0.91 0.73 0.85</td>
</tr>
<tr>
<td>Fulvic Acid</td>
<td>0.72 0.71 0.72 0.59</td>
<td>0.97 0.91 0.97 0.90</td>
<td>0.69 0.78 0.50 0.69</td>
</tr>
<tr>
<td>α-pinene</td>
<td>0.78 0.66 0.75 0.62</td>
<td>0.74 0.91 0.73 0.90</td>
<td>0.93 0.86 0.90 0.81</td>
</tr>
<tr>
<td>β-caryophyllene</td>
<td>0.82 0.71 0.80 0.65</td>
<td>0.71 0.93 0.69 0.92</td>
<td>0.94 0.86 0.92 0.81</td>
</tr>
<tr>
<td>linaloolSOA</td>
<td>0.77 0.62 0.75 0.56</td>
<td>0.68 0.83 0.66 0.81</td>
<td>0.87 0.74 0.81 0.64</td>
</tr>
<tr>
<td>α-terpinene</td>
<td>0.83 0.74 0.81 0.69</td>
<td>0.75 0.91 0.73 0.90</td>
<td>0.94 0.88 0.90 0.83</td>
</tr>
<tr>
<td>m-xylene</td>
<td>0.84 0.61 0.84 0.53</td>
<td>0.94 0.82 0.94 0.78</td>
<td>0.84 0.69 0.69 0.59</td>
</tr>
<tr>
<td>Diesel Exhaust, 0.25 hr</td>
<td>0.83 0.89 0.80 0.86</td>
<td>0.57 0.81 0.53 0.78</td>
<td>0.90 0.91 0.95 0.91</td>
</tr>
<tr>
<td>Diesel Exhaust, 2.25 hr</td>
<td>0.92 0.85 0.90 0.81</td>
<td>0.80 0.89 0.78 0.87</td>
<td>0.98 0.91 0.94 0.90</td>
</tr>
<tr>
<td>Diesel Exhaust, 4.25 hr</td>
<td>0.92 0.85 0.91 0.80</td>
<td>0.86 0.90 0.86 0.89</td>
<td>0.97 0.92 0.91 0.90</td>
</tr>
<tr>
<td>Ceanothus BBOA</td>
<td>0.88 0.85 0.85 0.80</td>
<td>0.70 0.84 0.68 0.81</td>
<td>0.94 0.90 0.94 0.89</td>
</tr>
<tr>
<td>Chamise BBOA</td>
<td>0.92 0.86 0.90 0.83</td>
<td>0.82 0.86 0.81 0.84</td>
<td>0.93 0.94 0.85 0.93</td>
</tr>
<tr>
<td>Palmetto BBOA</td>
<td>0.89 0.84 0.87 0.80</td>
<td>0.81 0.89 0.79 0.87</td>
<td>0.97 0.95 0.93 0.90</td>
</tr>
<tr>
<td>Juniper BBOA</td>
<td>0.92 0.87 0.90 0.83</td>
<td>0.84 0.85 0.83 0.82</td>
<td>0.97 0.94 0.94 0.94</td>
</tr>
<tr>
<td>Manzanita BBOA</td>
<td>0.88 0.85 0.85 0.81</td>
<td>0.72 0.85 0.70 0.83</td>
<td>0.95 0.91 0.92 0.94</td>
</tr>
<tr>
<td>Ponderosa Pine Duff BBOA</td>
<td>0.81 0.83 0.77 0.78</td>
<td>0.56 0.80 0.51 0.76</td>
<td>0.88 0.90 0.93 0.89</td>
</tr>
<tr>
<td>Ponderosa BBOA</td>
<td>0.85 0.85 0.83 0.81</td>
<td>0.64 0.82 0.61 0.78</td>
<td>0.93 0.91 0.91 0.88</td>
</tr>
<tr>
<td>Rice Straw BBOA</td>
<td>0.84 0.85 0.80 0.81</td>
<td>0.71 0.85 0.69 0.84</td>
<td>0.93 0.88 0.91 0.87</td>
</tr>
<tr>
<td>Sage and Rabbit Brush BBOA</td>
<td>0.88 0.86 0.87 0.82</td>
<td>0.93 0.89 0.93 0.87</td>
<td>0.91 0.94 0.79 0.92</td>
</tr>
<tr>
<td>Wax Myrtle BBOA</td>
<td>0.89 0.84 0.87 0.80</td>
<td>0.75 0.87 0.73 0.84</td>
<td>0.95 0.92 0.93 0.91</td>
</tr>
<tr>
<td>Levoglucosan</td>
<td>0.64 0.45 0.61 0.42</td>
<td>0.46 0.33 0.43 0.29</td>
<td>0.67 0.46 0.61 0.41</td>
</tr>
</tbody>
</table>
Table D.3. Correlations between input factors for the 3-factor synthetic cases.

<table>
<thead>
<tr>
<th></th>
<th>OOA-2</th>
<th>BBOA</th>
<th>FA</th>
<th>SOA</th>
<th>OOA-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$U^M$</td>
<td>$U^T$</td>
<td>$U^M$</td>
<td>$U^T$</td>
<td>$U^M$</td>
</tr>
<tr>
<td>HOA</td>
<td>0.81</td>
<td>0.84</td>
<td>0.88</td>
<td>0.84</td>
<td>0.39</td>
</tr>
<tr>
<td>OOA-1</td>
<td>0.84</td>
<td>0.55</td>
<td>0.81</td>
<td>0.55</td>
<td>0.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>OOA-2</th>
<th>BBOA</th>
<th>FA</th>
<th>SOA</th>
<th>OOA-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^M$</td>
<td>$R^T$</td>
<td>$R^M$</td>
<td>$R^T$</td>
<td>$R^M$</td>
</tr>
<tr>
<td>HOA</td>
<td>0.79</td>
<td>0.74</td>
<td>0.87</td>
<td>0.74</td>
<td>0.36</td>
</tr>
<tr>
<td>OOA-1</td>
<td>0.83</td>
<td>0.09</td>
<td>0.81</td>
<td>0.09</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Appendix E

Separation of Correlated Factors: Two-factor Synthetic Cases

Figure E.1a shows the relationship between the two PMF factors for each of the correlated variations on the base cases described in Sect. 2.2.3.2 and Fig. E.1b-d show how the solutions vary with FPEAK. The FPEAKs displayed represent an increase in Q/Q_{exp} of 0.2% to 0.6% from the FPEAK = 0 solution for each input case. As in the real data case, the behavior of positive values of FPEAK is to mix the time series, making them more correlated and the mass spectra less correlated, moving these solutions toward the upper left of the graph. Negative values of FPEAK do the opposite. The FPEAK = 0 solutions do not usually reproduce the input case; input cases are however reproduced closely with FPEAKs between -0.2 and +0.2. Note that, for most input cases, the PMF solutions produce a cluster of solutions with negative FPEAKs and a cluster with positive FPEAK solutions (in both cases with small dR/dFPEAK, i.e. the solutions have “hit a wall” and do not change much as FPEAK increases), and then a rapid transition between those groups of solutions (with large dR/dFPEAK). The amount of rotation between the factors, indicated by the length of the segments between marked FPEAKs, increases for input cases further from the origin. Cases with more correlated input factors (inputs closer to R = 1) rotate more with FPEAK while maintaining the same relative change in Q/Q_{exp} than cases with less correlated factors (inputs closer to R=0).

We are also interested in understanding how closely PMF solutions resemble the input cases, i.e., how well can PMF reproduce the input cases and does that depend on the correlation of input factors? We now consider the correlation between the input factors and the PMF output factors. In this case R_{MS} or R_{TS} = 1 means that the factor was retrieved perfectly by PMF. These
Figure E.1. Correlation of PMF factors in mass spectrum (R_{MS}) and time series (R_{TS}). Dots denote the correlation between the factors in an input case. Colored lines connect solutions of different FPEAKS for the same input case. Markers denote the FPEAK value of each solution. The FPEAK values displayed represent a change in Q of 0.2% from the FPEAK = 0 solution for each input case. a) Correlations between the factors for each input case. b) Cases that use the original time series but have varying correlations for the mass spectra. c) Cases that use the original mass spectra but have varying correlations for the time series. d) Cases that have varying correlations for both time series and mass spectra.
correlations are shown in Fig. E.2 for all of the cases described above. The range of FPEAKs
used is the same as those in Fig. E. 1. Factors with very low correlation in one axis and medium
correlation in the other are retrieved with little uncertainty, as expected. As either R_{MS} or R_{TS} of
the input factors increases towards 1 (cases further from the major axes origin), the factors have
increasing rotatability (a surrogate for uncertainty as described earlier). Cases in which both the
MS and TS are highly correlated (approaching the upper right of the plot) are usually retrieved
well at FPEAK = 0, but are retrieved poorly with only small changes in FPEAK. The effect is
continuous, but the deviations become especially large as R_{MS} and/or R_{TS} approach 0.8. PMF is
able to retrieve easily those factors whose contribution to the total data matrix X is very distinct
from that of the other factors and has difficulty separating highly correlated factors.
Figure E.2. Retrieval of input factors. Each set of internal axes corresponds to an input case and shows the correlation between the time series (vertical axes) and mass spectra (horizontal axes) of input factors and the PMF output factors. The outer axes reflect the relationship between the factors within each input case (same axes as Fig. E.1) as FPEAK is varied in each case. The internal axes are placed so that their intersection (1,1) is approximately at the point of correlation of the input case (the dots in Fig. E.1a). Correlations for HOA factors are shown in grey; correlations for OOA factors are shown in pink.
Appendix F

2-D Factorization of PToF m/z’s from Pittsburgh

This section briefly summarizes preliminary analysis of 2-D factorizations in PMF2 of the particle time-of-flight (PToF) data of individual m/z’s measured in Pittsburgh in 2002. Errors for the PToF data were calculated as described in Chapter 3 of this thesis for PToF data from ToF-AMS. All solutions shown below are for FPEAK = 0 unless noted otherwise. Size distributions are normalized so that their sum (not their area) equals 1.

F.1 Factorization of m/z 43 (C₂H₃O⁺ and C₃H₇⁺)

Fig. F.1. Q/Qexp vs. number of factors in solution (left) and Q/Qexp vs. FPEAK for solutions with 2 and 5 factors (right).

Fig. F.2. Fraction of variance and mass concentration for solutions with 1–6 factors. Solid black is residual mass not fit by the reconstruction.
Fig. F.3. Solution with 1 factor.

Fig. F.4. Solution with 2 factors.

Fig. F.5. Solution with 3 factors.

Fig. F.6. Solution with 4 factors.
Fig. F.7. Solution with 5 factors.

Fig. F.8. Solution with 6 factors.
F.2 Factorization of $m/z$ 43 ($\text{NO}_2^+$ with minor contribution from $\text{CH}_2\text{O}_2^+$)

**Fig. F.9.** $Q/Q_{\text{exp}}$ vs. number of factors in solution.

**Fig. F.10.** Fraction of variance and mass concentration for solutions with 1–6 factors. Solid black is residual mass not fit by the reconstruction.
Fig. F.11. Solution with 1 factor.

Fig. F.12. Solution with 2 factors.

Fig. F.13. Solution with 3 factors.

Fig. F.14. Solution with 4 factors.
**Fig. F.15.** Solution with 5 factors.

**Fig. F.16.** Solution with 6 factors.
F.3 Factorization of m/z 64 (SO$_2^+$ with minor contribution from C$_5$H$_4^+$)

Fig. F.17. Q/Qexp vs. number of factors in solution.

Fig. F.18. Fraction of variance and mass concentration for solutions with 1–6 factors. Solid black is residual mass not fit by the reconstruction.
Fig. F.19. Solution with 1 factor.

Fig. F.20. Solution with 2 factors.

Fig. F.21. Solution with 3 factors.

Fig. F.22. Solution with 4 factors.
Fig. F.23. Solution with 5 factors.

Fig. F.24. Solution with 6 factors.
Appendix G

Choice of solutions of the 3D factorizations

G.1 Choice of solution of the 3-vector model

In this section we explore solutions of the 3-vector model with two or more factors to choose the best solutions. Each family of solutions is examined against the criteria outlined in Sect. 3.3.4.3 to determine whether the factors are physically meaningful.

G.1.1 Solutions of the 3-vector model with two and three factors

The two-factor solution of the 3-vector model has only one family, in which one factor has a mass spectrum similar to OOA and the other has a mass spectrum that could be identified as HOA or BBOA. BBOA is generally differentiated from HOA in unit-mass-resolution data by BBOA’s higher contributions from characteristic biomass-burning markers at m/z’s 60 and 73. However, the contribution from these m/z’s in this mass spectrum is higher than in the HR-MS HOA mass spectrum but considerably lower than in the HR-MS BBOA mass spectrum. This factor appears to be a mix of the BBOA and HOA contributions that cannot be separated with only two factors. Thus we consider solutions of the 3-vector model with three factors.

The three-factor solutions of the 3-vector model fall into two families. The family with the lower $Q/Q_{exp}$ values has HOA, BBOA, and OOA factors. These factors appear to be physically meaningful; thus this solution could be acceptable. However, the solution lacks an LOA factor.
The largest contribution to $Q/Q_{exp}$ for this solution comes from $m/z$ 58, which in the HR-MS solution was composed almost completely of organic-nitrogen fragments (C$_3$H$_8$N$^+$ and C$_2$H$_4$NO$^-$). Though LOA has a small contribution to the total aerosol mass, its chemical composition is distinctive, and we would expect to find such a factor in the PToF data.

The other 3-vector family with three factors has a $Q/Q_{exp}$ that is 0.9% higher than the first family. We note that this difference is much larger than the final convergence criterion used to compute the solutions (0.001% $Q/Q_{exp}$), suggesting that the families indeed represent different local minima in the solution space. We do not know whether this increase in $Q/Q_{exp}$ is meaningful (i.e., large enough to imply that the solution is bad) because comparisons between families of solutions have not been discussed in the literature of 3D aerosol factorization. The closest literature comparison is the 2D factorization of an AMS dataset by DeCarlo et al. (2010), in which the $Q/Q_{exp}$ values for 50 seed solutions with four factors had a range of 1.5% above the minimum $Q/Q_{exp}$. In this case, the family with the best solution included both the lowest and highest $Q/Q_{exp}$ values. We might therefore expect a 0.9% increase in $Q/Q_{exp}$ to be acceptable; however, it is not clear that the properties of 2D and 3D factorizations of different datasets should be similar. Therefore we make no judgment in the present case about the meaning of increases of $Q/Q_{exp}$.

The factors in the higher $Q/Q_{exp}$ family are identified as OOA, BBOA, and LOA. The OOA and BBOA factors have mass spectra, time series, and size distribution that are similar to those of the lower $Q/Q_{exp}$ family. The LOA, in contrast, is more difficult to recognize because its mass spectrum is distorted compared to the HR-MS LOA factor. The HR-MS LOA mass spectrum has significant contributions from $m/z$ ‘s also prominent in the HOA and BBOA spectra. These
peaks, at \( m/z \) 27 and 29, 41 and 43, 55 and 57, etc., form a characteristic “picket fence” pattern of peaks containing \( C_nH_{2n-1}^+ \) and \( C_nH_{2n-1}CO^+ \) ions at the lower \( m/z \) of each pair and \( C_nH_{2n+1}^+ \) and \( C_nH_{2n+1}CO^+ \) ions at the higher \( m/z \) of each pair (McLafferty and Turecek, 1993). However, these characteristic hydrocarbon peaks are much less prominent in the LOA mass spectrum found in this study. The HR-MS spectrum contains another characteristic peak at \( m/z \) 91, which is present in our LOA spectrum at similarly higher abundance relative to the neighboring peaks. Specifically, the HR-MS spectrum has a 3% contribution from \( m/z \) 91, which is \( \sim 10 \) times higher than the adjacent peaks in the spectrum, while our LOA has a 4.5% contribution from \( m/z \) 91, which is \( \sim 12 \) times higher than the adjacent peaks. The major difference between the HR-MS LOA spectrum and ours is at \( m/z \) 58. The nitrogen-containing fragments at \( m/z \) 58 contributed only 1% of the HR-MS LOA spectrum, but \( m/z \) 58 contributes 19% of the signal to LOA in this solution. The unusual appearance of our LOA spectrum would lead us to reject this factor as not physically meaningful if we did not know of its existence from the HR-MS solution. Nevertheless, the existence of the LOA factor in this solution demonstrates that LOA is present in the PToF dataset, even though the mass spectrum is different than that in the HR-MS factor.

G.1.2 Solutions of the 3-vector model with four factors

Since all of the HR-MS factors are found in the three-factor solutions, we expect to find solutions of the PToF dataset with four or more factors that include all four of the HR-MS factors. Such a solution with four factors is found, but not in the family with the lowest \( Q/Q_{exp} \) value. The four-factor solutions are discussed in order of increasing \( Q/Q_{exp} \) value.

The four-factor solutions can be grouped into four families (Fig. H6). The family with the lowest \( Q/Q_{exp} \) values has two recognizable factors and two factors that do not appear to have
physically meaningful mass spectra (Fig. H7). The recognizable factors represent HOA and BBOA and have mass spectra, size distribution, and time series similar to the three-factor solution with the lowest $Q/Q_{exp}$ value. The mass spectra of the other factors each have large contributions from $m/z$’s that are not usually so prominent. One factor has a mass spectrum dominated by signals at $m/z$ 44 and related peaks, as defined in the fragmentation table ($m/z$’s 28, 18, 17, and 16). The fraction of signal contributed by $m/z$ 44 ($f_{44}$) is proportional to the aerosol oxygen-to-carbon ratio (O:C), which is also proportional to aerosol age, so high $f_{44}$ represents highly oxidized, aged aerosol (Aiken et al., 2008). This factor might therefore represent OOA, and in fact has a high correlation with the OOA factor from the three-factor solution ($R = 0.86$). But this high-$m/z$-44 factor contributes only half of the mass of the OOA factor in the three-factor solution. Curiously, this mass spectrum has $f_{44} = 0.39$, and the group of related peaks contributes 90% of the signal in this mass spectrum. This spectrum represents aerosol with an elemental oxygen-to-carbon ratio (O:C) of 1.6 (Aiken et al., 2008) — much higher than O:C measured in ambient or laboratory aerosol by the AMS (Ng et al., 2010). O:C can be combined with the estimated hydrogen-to-carbon ratio (H:C, Ng et al., 2011a) to estimate the average carbon oxidation state ($\overline{OS_C}$) of the aerosol from this factor (Kroll et al., 2011). This factor has $\overline{OS_C} = +2.0$. However, the $\overline{OS_C}$ of atmospheric aerosols rarely exceeds +1 because compounds with higher oxidation states tend to fragment and/or evaporate (Kroll et al., 2011). Thus the factor in this solution with $f_{44} = 0.39$ improves the fit of the factorization, as evidenced by its appearance in the family of solutions with the lowest $Q/Q_{exp}$, but it is not physically meaningful.

The other unrealistic factor in this solution has a mass spectrum that resembles HOA but is dominated by $m/z$ 43. Compared to the HR-MS HOA, the mass spectrum of this factor has more signal at many $m/z$’s ≤ 45 and less signal at most $m/z$’s > 45. Despite our factor’s mass spectral
HOA features, its time series is negatively correlated with that of HOA in the three-factor solution (R = -0.12). This factor’s time series is actually correlated with the time series of OOA in the three-factor solution (R = 0.77) and accounts for 75% of the OOA mass from the three-factor solution. Yet, the mass spectrum has no contribution from m/z 44, our most important marker for OOA. Thus the evidence for identifying this factor as HOA or OOA is contradictory.

The factors dominated by m/z 43 and 44 likely represent a “splitting” of the OOA factor, i.e., the OOA factor has been divided into two factors (Ulbrich et al., 2009). This division can be demonstrated by summing the time series from the factors dominated by m/z 43 and m/z 44. The summed time series has very strong correlation with the time series of OOA in the three-factor solution (R=0.95) and similar mass contribution. The factor splitting can further be demonstrated by using the average mass contribution from these two factors as weights to calculate the weighted-average mass spectrum from these two factors. This weighted average mass spectrum has a strong correlation with the HR-MS OOA mass spectrum (R=0.98). From this evidence, we conclude that the factors dominated by m/z 43 and m/z 44 are not individually physically meaningful, and so we reject this family of solutions.

The next family of four-factor solutions has Q/Q_{exp} only 0.1% higher than the first family, but these factors are all recognizable aerosol types (Fig. 3.4). The four factors can be identified as OOA, HOA, BBOA, and LOA. The LOA mass spectrum has even higher fractions of m/z 58 and m/z 91 than in the three-factor solutions; however, these increases actually reflect a decreased contribution from the characteristic hydrocarbon m/z’s that also contribute to HOA and BBOA. The largest contribution to Q/Q_{exp} for this family is now from m/z 15, and the remaining structure in the Q/Q_{exp} time series has only small features that have little correlation
with factor time series or the total mass time series. All of the factors in this family of solutions are physically meaningful; thus, this solution could be acceptable.

The next family has $Q/Q_{exp}$ 0.4% higher than the first family. The four factors in this family can be recognized from their mass spectra as OOA, HOA, and two factors representing BBOA (Fig. H.8). The occurrence of two factors with the same aerosol type identification is rare in the solutions of the 3-vector model (Table 3.2), so the occurrence of two BBOA factors is intriguing. The two BBOA factors have quite different size distributions. One factor has a somewhat narrow size distribution with particle $d_{wa}$ that ranges ~40 nm to ~400 nm and has a mode at $d_{wa}$ ~180 nm. The other BBOA factor has a broader distribution with $d_{wa}$ that extends to 1150 nm and has a mode at a larger $d_{wa}$ near 370 nm. These two BBOA factors have very similar mass spectra (R=0.94). The presence of two BBOA factors suggests that the BBOA factor has been split. Indeed, the sum of their time series has very high correlation with the time series of the BBOA factor in the previous family discussed (R=0.98) and has 21% more mass. In addition, the time series of these two factors are moderately correlated (R=0.65), a characteristic of split factors observed in previous studies (Ulbrich et al., 2009). The split of the BBOA factor is most likely because the 3-vector model is not able to fit the BBOA size distribution variability, which might be better fit in the vector-matrix model.

Another interesting feature of this solution is the broad size distribution of the HOA factor. The HOA factor has a bimodal size distribution that spans the entire diameter range (20 – 1300 nm) with modes at $d_{wa}$ of ~115 nm and ~410 nm (Fig. H.8). The broad size distribution of this HOA factor is similar to the average size distribution of HOA in Pittsburgh aerosol (estimated as 12.3 * m/z 57 - 0.02 * m/z 44, Zhang et al., 2005a; Zhang et al., 2005b). However, the size
distribution of the Pittsburgh HOA is flatter, i.e., distinct modes are not easily identified. Since the other factors in this solution are physically meaningful, this solution may be acceptable. The only failing of this solution appears to be that it is missing an LOA factor, which is reflected by the high $Q/Q_{exp}$ contribution from $m/z$ 58.

The last four-factor family has $Q/Q_{exp}$ 0.8% higher than the first family. This family has factors with mass spectra that are recognized as OOA, HOA, and BBOA. The fourth factor is dominated by $m/z$ 15; in factor, $m/z$ 15 contributes 70% of the signal in the mass spectrum of this factor. The remaining signal is contributed by the characteristic hydrocarbon $m/z$’s common to the HOA and BBOA spectra. Because of the dominance of $m/z$ 15 in this factor, no other factor mass spectra in this solution have a contribution from $m/z$ 15. The size distribution of this factor includes signal at nominal particle sizes too small to actually have particle signal. The signal at these small particle sizes is likely from gas-phase interference of $^{15}$N$^+$ whose influence was not removed perfectly with the fragmentation matrix or downweighted like $N^{15}$N and CO$_2$. The features of the factor dominated by $m/z$ 15 demonstrate that it is not physically meaningful; thus we reject this solution.

In summary, the four-factor solutions of the 3-vector model included two families with potentially acceptable solutions. One of these families included all four of the HR-MS factors, while the other had two instances of the BBOA factor and lacked a factor for LOA. We now explore solutions with five or more factors to determine whether they might contain additional information, e.g., the four HR-MS factors and a new factor.
G.1.3 Solutions of the 3-vector model with five or more factors

The five-factor solutions fall into five families (Fig. H.6), and all contain factors that the fail criterion of being physically meaningful. The four families with the lowest $Q/Q_{exp}$ values have the same factors as in the four-factor families with the lowest $Q/Q_{exp}$, except that OOA is never present as one factor (Table 3.2). Instead, all of these families have the $m/z$ 43- and $m/z$ 44-dominated factors that represent a splitting of the OOA factor. In contrast, the five-factor family with the highest $Q/Q_{exp}$ does have a single OOA factor, but it also includes the factor dominated by $m/z$ 15 already deemed not physically meaningful. We therefore reject all of the five-factor solutions.

The solutions of the 3-vector model with six factors fall into seven families. In all of these families, OOA is split into the factors dominated by $m/z$ 43 and $m/z$ 44. The solution with the lowest $Q/Q_{exp}$ value also includes factors representing HOA, BBOA, and LOA, and the factor dominated by $m/z$ 15. Many of the families with lower $Q/Q_{exp}$ values have multiple BBOA factors, while some of the families with higher $Q/Q_{exp}$ values have multiple HOA factors. No families have multiple instances of both BBOA and HOA factors. We reject these solutions because they do not containing physically meaningful factors, or because their factors do not provide new information compared to solutions with fewer factors.

The pattern of factors without additional physical meaningfulness continues in solutions of the 3-vector model with seven or eight factors. These solutions have a higher-dimensional solution space, which has more possibilities for local minima. Accordingly, we observe more families than in the solutions with fewer factors, but these solutions include the same types of factors.
Multiple occurrences of some factors are more common in the solutions with seven and eight factors. However, no new factor types are observed in any of these solutions.

In summary, the most acceptable solutions of the 3-vector model had four factors. One of the four-factor solutions included all four of the factors in the HR-MS solution. The other family included all of the HR-MS factors except LOA, but instead had two occurrences of the BBOA factor with similar mass spectra and time series but different size distributions. Because we do not have sufficient support for the validity of two BBOA factors, we choose the family that had the four factors in the HR-MS solution as the best solution of the 3-vector model.

G.2 Choice of the number of factors in the solution of the constrained vector-matrix model

In this section we explore solutions of the fully constrained ($\beta = 0$) vector-matrix model with four or more factors to determine whether new, physically meaningful factors can be identified. Each family of solutions is examined against the criteria outlined in Sect. 3.3.4.3 to determine whether the factors are physically meaningful.

The four-factor solutions of the constrained vector-matrix model form one family. The four factors are the a priori factors specified for this model and are described in Sect. 3.4.2.

The five-factor solutions of the constrained vector-matrix model form only one family. In these solutions, the fifth factor resembles HOA, but has a greater contribution from $m/z$ 43, and the characteristic “picket fence” pattern of HOA is shifted to $m/z's \geq 43$ (Fig. H.9). This HOA-like spectrum may not be physically meaningful because it has only 6.5% of its signal in $m/z's < 43$, compared to 33% for the HR-MS HOA spectrum and 30% for a “standard” HOA spectrum derived from PMF analyses of fifteen urban AMS datasets (Ng et al., 2011b). Spectra with such
a small fraction of signal in the low-mass fragments are not found in the AMS Spectral Database (Ulbrich et al., 2011) or in other electron impact spectra (McLafferty and Turecek, 1993). In addition, there is evidence that this factor is a splitting of the HOA factor. In fact, the sum of the size-distribution–time-series matrices from this factor and the HOA factor from this solution is very similar (R = 0.99) to the size-distribution–time-series matrix of HOA in the four-factor solution. Thus we conclude that this factor has split the HOA and does not constitute a new, useful factor.

The six-factor solution has two families. The family with the lower $Q/Q_{exp}$ values has the four constrained factors, the split HOA factor found in the five-factor solution, and a factor dominated by $m/z$ 43. This $m/z$-43Dominated factor is mainly a split of the OOA factor, as was observed in the 3-vector solutions. This factor also takes some mass from BBOA during the large BBOA event on 21 March. From this evidence we conclude that this factor is not physically meaningful. We therefore reject this family of solutions. In the other family of six-factor solutions, the $m/z$ 43 spectrum is replaced by a spectrum dominated by $m/z$ 44. The factor dominated by $m/z$ 44 is also a split of the OOA factor, and its spectrum has even higher $f_{44}$ (43%) than the spectrum dominated by $m/z$ 44 in the 3-vector solutions ($f_{44} = 40\%$) that was rejected as nonphysical. Thus we conclude that this factor also fails the criterion of being physically meaningful. We therefore reject the six-factor solutions. The seven- and eight-factor solutions have more families, but also contain factors with unusual mass spectra that are not physically meaningful. These solutions are also rejected.
Thus, only the four-factor solution has factors that are all physically meaningful. Therefore we choose the four-factor solution and explore the effect of increasing $\beta$ to relax the constraint on the a priori spectra, as discussed in Sect.3.4.2.1.
Appendix H

Supporting Information for Chapter 3

H.1 Literature Review: 2-dimensional factorization of datasets that include size-distribution data

In this section, we discuss and summarize research that reports the application of mathematical techniques to datasets that include particle size information. We first present an overview of the mathematical techniques used in studies of 2-dimensional (2D) datasets. Then we briefly summarize the arrangement of datasets containing particle size information into 2D matrices. The datasets include particle size information in one of two ways: (1) the dataset contains aerosol size distributions, and may also include simultaneous measurements of gas-phase and/or bulk aerosol-phase chemical constituents; or (2) the dataset contains size-resolved aerosol chemical composition, and the factors obtained from analysis of different size ranges were compared. Finally, we summarize the studies in the literature that have applied factorization methods to these datasets.

H.1.1 Mathematical techniques for 2-dimensional factor analysis

Various mathematical techniques have been applied to datasets that include size distribution or size-resolved aerosol composition information. Collectively, the goal of these techniques is the same: to determine particle sources. Several terms have been used in the literature to describe these related techniques, including factor analysis, source apportionment, and matrix
factorization. Each of these terms has a different technical definition and common usage.

Although “factor analysis” technically refers to multivariate analyses that produce orthogonal factors (Malinowski, 1991), it is commonly used to refer to a variety of multivariate analyses, regardless of the orthogonality of the factors. In this work, we accept the common usage and refer to the class of multivariate analyses that have been applied the datasets of interest as “factor analysis.”

We now discuss how the 2D factorization model represents aerosol processes. A 2D matrix is deconvolved such that each factor is composed of two one-dimensional (1-D) vectors (i.e., a bilinear unmixing model). Several examples of this scheme are shown in Fig. H.10. One factor vector describes how much of that factor is present in each sample and is called a time series. The other vector describes the composition of the aerosol in that factor and is called a profile. A factor’s profile shows the fractional contribution of the constituents contained in the rows of the input matrix. For example, in the matrix shown in Fig. H.10a, each matrix row contains the concentrations of particles at a given size, and each factor profile also contains a size distribution. In contrast, the matrix rows in Fig. H.10b contain concentrations of particles at given sizes and simultaneously measured concentrations of gas-phase species. The factor profiles from this matrix thus contain a size distribution and relative concentrations of the gas-phase species.

The data matrix is reconstructed by the linear combination of some number of factors, and each measured sample (matrix rows) can have contributions from more than one factor. This reconstruction is described by

\[ x_{ij} = \sum_p a_{ip} c_{pj} + e_{ij}, \]  

(H.1)
where $i$ and $j$ are the row and column indices for the matrix, respectively; $p$ is the number of factors; $x_{ij}$ is an element of the $m \times n$ data matrix $X$ to be factored; $a_{ip}$ is an element of the $m \times p$ matrix $A$, the columns of which contain the factor time series; $c_{pj}$ is an element of the $p \times n$ matrix $C$, the rows of which contain the factor profiles; and $e_{ij}$ is an element of the $m \times n$ matrix $E$ of the residuals of the solution, i.e., the difference between the measured data and the reconstruction.

Several factor analytical methods are available to solve the bilinear unmixing model described in Eq. (H.1). The methods differ by the requirements of a priori knowledge about the factor profiles and constraints placed on the factor profiles and time series. In principal component analysis (PCA, Malinowski, 1991) and absolute principal component analysis (APCA, Thurston and Spengler, 1985), no a priori knowledge about the profiles is required. In these methods, the factor profiles are orthogonal to each other, and the factor time series are also orthogonal to each other. The orthogonality requirement creates factor profiles and time series with both positive and negative values. In contrast, Positive Matrix Factorization (PMF, Paatero, 1997) does not require a priori knowledge about the profiles. Furthermore, orthogonal factors are not possible in PMF2 as the elements of the factor profiles and time series are constrained to be positive. The positivity requirement means that all profiles will have positive signals, and the time series will have positive mass. Consequently, PMF factors are usually more physically meaningful than those from PCA. In contrast to these two methods, the factor profiles are completely prescribed in the Chemical Mass Balance approach (CMB, Friedlander, 1973), and no constraints are placed on the factor time series.
Instead of using factor analytical methods, samples can be grouped into clusters based on sample similarity. Cluster analysis groups measurements with similar characteristics, and these clusters can be associated with sources of measured aerosol (Murphy et al., 2003; Dillner et al., 2005; Marcolli et al., 2006; Beddows et al., 2009). The clusters can also be represented with two 1-D vectors: the cluster profile is the average composition of the samples in the cluster, and the time series is the number of samples in the cluster over average sampling periods. In most cluster analyses, an entire sample vector is assigned to a group, and therefore each total measurement can be assigned to only one cluster. A notable exception is fuzzy cluster analysis, in which sample vectors can be assigned to multiple clusters with an associated degree of membership to each group (Bezdek et al., 1981).

These mathematical techniques have been applied to 2D datasets that include particle size information. We now summarize these studies in order of increasing complexity of the data in the factorization matrix.

### H.1.2 Arrangements of datasets including particle size information for factor analysis

The datasets that include size distributions and whose factorization has been reported in the literature can be divided into two broad categories. In the first category, samples include one set of measurements at each time. These measurements may include any of the following: particle size distributions, particle-phase bulk chemical composition, or gas-phase chemical composition. The measurements are arranged as 2D matrix in which one dimension contains the measured data for each time step and the other dimension is time. This category of datasets contains four subcategories, depending on the type of measured data (Fig. H.10a–d). All four subcategories include particle size distributions, i.e., the number or volume concentration of particles at
multiple particle sizes. One subcategory uses only size distributions in the input matrix (Fig. H.10a), while the other subcategories couple size distributions and different combinations of gas-phase composition and bulk aerosol chemical composition data (Fig. H.10b–d).

In the second category, samples include two dependent measurements at each time: particle size, and the particles’ chemical composition at each size, i.e., size-resolved composition data. The size-resolved composition data can be arranged in two forms for factorization (Fig. H.10e–f). In the first arrangement, the chemical composition data from each particle size form a separate 2D matrix, and each 2D matrix is factored independently of the others (Fig. H.10e). In this arrangement, a single source may contribute to different size ranges, but is not required to have the same chemical composition at every size. In contrast, the size-resolved composition data can be arranged such that the measured concentrations from different particles are appended as rows of a single 2D matrix (Fig. H.10f). In this arrangement, the chemical composition of each factor must be the same for all particle sizes.

S1.3 Research reporting 2-dimensional factorizations using particle-size information

Research that has reported 2D factor analysis of aerosol number or mass size distributions, or datasets of size-resolved, aerosol chemical composition are summarized briefly in Table H.2 and in greater detail in Tables H.3–H.8. Schematics of the factorization of six dataset types are shown in Fig. H.10. Each of the six categories is discussed below.

The first category of studies applies factor analysis techniques to datasets of aerosol number distributions or combined aerosol number and mass distributions (Fig. H.10a, Table H.3). A good example from this category is the study by Costabile et al. (2009), who used PCA to
analyze aerosol number distributions collected over two years from eight locations in and near Leipzig, Germany. The authors analyzed the data in two ways: (1) They factored the measurements from each site separately and (2) they factored combined measurements from selected subsets of the locations. Combining measurements from different locations allowed the authors to assess the temporal and spatial variation in the number size distribution. The authors identify factors in the nucleation mode, Aitken mode, and accumulation mode size ranges. The factors have names such as “fresh, roadside,” “fresh, background,” “rural,” “urban traffic”, and “continental.” The authors assigned some of their factors to sources, but noted that some aerosol modes have contributions from multiple sources, e.g., long-range-transported and primary urban aerosol. Another study in this category acknowledges that particle size distribution components may not be directly related to particle sources (Chan and Mozurkewich, 2007a), and so they combine simplified representations of size distributions derived by APCA from three rural and urban locations near Toronto with gas-phase concentrations of CO, NO$_x$, SO$_2$, and O$_x$ (NO$_2$ + O$_3$) and wind speed to associate the different aerosol modes with photochemical processing, regional pollution, boundary layer dynamics, local anthropogenic emissions, and processed nucleated particles (Chan and Mozurkewich, 2007b). However, most of the other studies in this category assign factors to sources such as stationary combustion sources, local traffic, spark-ignition gasoline emissions, and secondary aerosol, but only rarely note that such assignments can only be tentative without aerosol composition data (Kim et al., 2004). In addition to identifying aerosol modes, Costabile et al. (2009) presented a paradigm for transformation of the aerosol size distribution from the local to the regional scale, showing that aerosol near sources has high temporal and spatial variability, but that the size distributions become similar in both space and time as aerosol is aged and transported regionally.
The second category of studies pairs aerosol number distributions with the concentrations of simultaneously measured gas-phase species (Fig. H.10b, Table H.4). In one such study, Thimmaiah et al. (2009) propose the factorization of size distribution and gas-phase data as a cost-effective method for informing air-quality management programs. However, only four factors are identified in this PMF study (ozone-rich, transported ozone/ozone precursors; NO\textsubscript{x}-rich diesel emissions; traffic-spark ignition vehicles; and local heating sources), and these factors are weakly supported by correlations between the factor time series and the time series of the species included in the factorization. Because the gas-phase species are important contributors to the factors themselves, it is not surprising that the correlations are high. For example, a factor attributed to transported ozone and ozone precursors has a strong correlation with ozone concentrations, and only this factor has an appreciable contribution from ozone. Stronger support for factor identification requires correlations with tracers external to the factor analysis (Zhang et al., 2005; Lanz et al., 2007). The Thimmaiah et al. study does not convincingly hold up its claim for producing useful information. In the only other published study in this category, Wahlin et al. (2001) measured aerosol number distributions, CO, and NO\textsubscript{x} at roadside-sampling locations, and applied PCA to these data with the goal of separating diesel and gasoline contributions to ultrafine particles. The diesel and gasoline contributions could not be completely separated, and one factor was attributed to diesel vehicles, while another was attributed to both gasoline and diesel vehicles. The authors concluded that more reliable separation of the gasoline and diesel contributions requires a specific tracer for diesel emissions; furthermore, they noted that better understanding of particle sources and transformation could come from combining particle size distributions and particle composition data.
The third category of studies pairs aerosol number distributions with chemically speciated measurements of the bulk aerosol (Fig. H.10c, Table H.5). Such datasets must be constructed carefully with respect to the particle size distribution used in the analysis. The aerosol number distribution is dominated by very small particles, but the aerosol chemical composition comes mainly from the large particles that contribute most of the particle mass (Seinfeld and Pandis, 1998). However, the particle mass and volume distributions usually agree well, and so including these together should be satisfactory for factor analysis. To address the discrepancy and differences in the uncertainties between the number size distribution and composition data, Larson et al. (2006) decreased the weight of the size information by a factor of 10 in the factorization of measured particle number distributions, volume distributions, and chemically resolved species from PM$_{2.5}$ (particulate matter with diameter ≤ 2.5 µm). In this study, the degree of weighting of the size information had only a minor effect on the factors’ size distributions and average mass contribution, but the effect on the factors’ chemical composition was not described. Factors identified in this study include vegetative burning, aged sea salt, and metals processing.

The final category of datasets that include non-chemically resolved size-distribution data combines them with both gas-phase and bulk particle composition data (Fig. H.10d, Table H.6). Like the research that combined particle size distribution and chemical composition data, in these studies the input aerosol size distributions usually encompassed a smaller size range than the bulk composition data. For example, Zhou et al. (2005a) applied PMF to a dataset that combined aerosol number distributions, concentrations of particulate nitrate, sulfate, and 11 trace metals, and concentrations of five gas-phase species. Factor names were assigned based mainly on the chemical composition of the factors, which included two secondary nitrate factors, coal-fired
power plant, steel mill, and nucleation. In this study, sulfate (and to a lesser degree, nitrate) had substantial contributions to the aerosol composition of all factors. Although each factor included a size distribution, little information was given about how well the overall size distributions were fit, and it was unclear how strongly the size distribution information influenced the factorization results.

We have now reviewed the studies that include particle size distributions in the factorization matrix. In general, the particle size information may have been too strongly linked to sources when little or no particle chemical composition was included, or undervalued when the analysis relied predominantly on bulk composition data for source identification. We now consider studies that use size-resolved aerosol composition datasets to explore the sources of ambient aerosol.

The remaining 2D factorization studies explore such size-resolved chemical composition datasets. In most of these studies, factor analysis is performed separately on the chemical species measured for each size range (Fig. H.10e, Table H.7). The studies that factor the composition from each size range separately fall into two main groups based on the size resolution of the data: (1) studies that use species from overlapping bulk PM sizes (e.g., PM$_{2.5}$, PM$_{10}$) and (2) those that use multistage samplers to separate particles by size. Two studies of the latter type are reviewed here. Kleeman et al. (2009) measured 8 molecular organic tracers from six stages of a MOUDI sampler (0.055–1.8 µm) at urban locations in California. This study used a custom source-apportionment algorithm to relate tracers for five sources (wood burning, meat cooking, motor oil, gasoline, and diesel fuel) to measured elemental and organic carbon concentrations. The dominant sources of elemental carbon were found to be gasoline and diesel
vehicle exhaust. Little organic carbon mass was attributed to vehicle exhaust. The dominant sources of organic carbon were found to be wood burning and meat cooking. Organic carbon that could not be assigned to the selected sources was 10–24% of the PM$_{1.8}$ mass and 0–58% of the PM$_{0.1}$ mass and may come from oil and gas refining in the region or from SOA. In another study, Han et al. (2006) analyzed a dataset of 19 elements measured from an 8-stage DRUM sampler (0.07 to ~12 µm) collected at Gosan, Korea, in 2002. PMF was applied separately to the chemical composition data from each stage, i.e., particle size range. The authors identified fifteen sources in total, with four to eight sources contributing to any single size range. Most sources had a strong size dependence. For example, local soil and sea salt were present mainly in coarse sizes, while coal combustion particles and diesel vehicle particles were identified in only the fine and ultrafine sizes, respectively. The chemical profiles identified for the same source in different size ranges were shown to have high similarity.

The final type 2D factorization is based on the assumption that particles from a single source have the same composition across a broad size range. Instead of factoring size-resolved composition data separately for each particle size range, Amato et al. (2009) combined speciated chemical measurements from different overlapping size ranges (PM$_{10}$, PM$_{2.5}$, and PM$_{1}$) into one factorization matrix (Fig. H.10f, Table H.8). This dataset could not be factored using a 3-dimensional (3D) model because the measurements from different sizes were not collected simultaneously. Instead, the authors arranged their input matrix such that species from each size were appended as rows of the matrix. In this matrix arrangement, the time series of each factor includes the contribution of that factor at each size, and each factor must have the same chemical profile for all sizes. The assumption that the factors have the same chemical profile for all sizes was not tested directly because the data from each size were not factored separately. However,
the contributions of the sources to each particle size matched previous knowledge about these sources. For example, aged sea salt was found with high concentrations in PM$_{10}$, lower concentrations in PM$_{2.5}$, and negligible mass contributions to PM$_{1}$. In contrast, vehicle exhaust had the comparable contributions to all three size ranges.

In summary, multiple 2D factorization approaches have been applied to aerosol datasets that include particle size distributions. These studies attempt to understand the processes influencing ambient particle size distributions, but at best provide speculative assignments to sources in the absence of chemical information. The addition of some chemical information from gas-phase species gave insight to particle sources only in one carefully constructed study. Combining aerosol size distributions with particle composition data enabled more complete attempts to characterize the size distribution of aerosol sources. The 2D factorization approaches that used size-resolved aerosol composition are the most promising of the 2D cases for obtaining a size-resolved source apportionment. However, these studies have coarse time and size resolution, and are unable to address the dynamic nature of the aerosol size distribution as it evolves through atmospheric processes.
H.2 m/z’s with organic signal omitted from the present study

The standard fragmentation matrix for organics does not assign organic signal to some m/z’s that are known to have organic signal but also have interferences from air or inorganic ion signals (m/z 14, 32, 33, 36, 39, 40, 46, and 47). This omission removes only a small part of the organic aerosol mass. Of the omitted m/z’s, the greatest contribution to the organic mass comes from m/z’s 39 and 40. Each of these two m/z’s contribute ~2% of the total organic mass of the MS mode data, and the omitted masses in total contribute 7% of the total mass (Aiken et al., 2009).

In addition to the masses normally omitted by the fragmentation matrix, we also omit m/z’s 12 and 30 from this analysis. These m/z’s can be examined in the high-resolution MS-mode data (Aiken et al., 2010) and appear to deviate from the assumptions of the standard fragmentation matrix. m/z 12 has contributions from an unidentified ion that can contribute up to 25% of the signal at this m/z, and we do not know how this unidentified ion might be represented in the size distribution. At m/z 30, the organic signal is predicted to come from the $^{13}$C isotope of C$_2$H$_5^+$, but is instead mainly from CH$_2$O$^+$ ions, for which a fragmentation ratio to another peak has not been characterized.
H.3 Comparison of the PMF3 and ME-2 algorithms for solving the 3-vector model

The 3-vector model can be solved by PMF3 or ME-2, which use different convergence algorithms to minimize $Q$. Both algorithms iteratively minimize $Q$, and each step in the iteration tests a possible solution in the “$Q$ space”. The $Q$ space is a function of the data and error matrices and the model used to fit the data, but is independent of the algorithm used to explore that space. We expect that both algorithms have landed in the same local minimum when they find solutions with the same factors and similar $Q$ values.

ME-2 minimizes $Q$ using the conjugate gradient algorithm (Paatero, 1999), which stops after encountering a sharp change in the $Q$ gradient. Such a change marks the transition from a “wall” to the “floor” of a local minimum in the $Q$ space. In contrast, PMF3 uses the Gauss-Newton algorithm (Paatero, 1999), which seeks the lowest point in the local minimum in the $Q$ space. Logically, the lowest point in the minimum could be lower than the edge of the floor. This result is observed in solutions of the 3-vector model in this study (Fig. 3.3b). These solutions have downweighted $Q/Q_{exp}$ values with differences similar to those between solutions in the same family solved by the same algorithm ($< 10^{-4} Q/Q_{exp}$ units). However, the unweighted $Q/Q_{exp}$ values show somewhat larger differences (Fig. 3.3a). The difference in the unweighted $Q/Q_{exp}$ values between the solutions indicates that the fits differ somewhat, and that differences in the fits occur mainly in the downweighted $m/z$-size combinations. In fact, comparing the best four-factor solution from each algorithm shows that the main difference in the fits is at $m/z$ 100, where the PMF3 solution has a lower $Q$ contribution. Since $m/z$ 100 has low SNR at all sizes, the difference between the Q contributions is amplified when the downweighting is removed.
We do not believe that the lower $Q/Q_{exp}$ values from PMF3 for this dataset imply that the PMF3 solution is better, or that PMF3 uses a better algorithm. Furthermore, the Gauss-Newton method used by PMF3 is less efficient than the conjugate-gradient algorithm used by ME-2 when solving large problems (Paatero, 1999). We observed this difference in the speed of computing solutions of the 3-vector model with six or more factors (Fig. H.11). Therefore, we suggest using ME-2 to solve the 3-vector model for its speed advantage, and also when researchers plan to compare to other models (which cannot be computed by PMF3).

S4 Solutions of the unconstrained vector-matrix model

Compared to the 3-vector solutions, solutions of the unconstrained vector-matrix model with the same number of factors have lower $Q/Q_{exp}$ values (Fig. 3.3). The low $Q/Q_{exp}$ values for the vector-matrix solutions imply that the vector-matrix model can fit more of the data than the 3-vector model. In other words, the two vectors in the 3-vector model fit the size-time information less well than the matrix in the vector-matrix model. The better fit of the vector-matrix model is consistent with this model’s greater degrees of freedom [Eqns. (3.6) and (3.7)]. However, the large number of degrees of freedom in the vector-matrix model also allows the model to fit a large amount of noise. The ability to fit noise in this low SNR dataset distorts the mass spectra in the model solutions.

No solutions of the vector-matrix model have mass spectra that are all physically meaningful (Table 3.3). For example, in solutions with at least three factors, a single factor for OOA is obtained in only one solution. This solution contains OOA, HOA, and LOA factors, but not a BBOA factor. The omission of BBOA is surprising since this factor has a larger mass fraction than LOA when both are found in the same solution. In addition, this solution’s increased $Q/Q_{exp}$
may indicate a poor fit. However, in all other unconstrained vector-matrix solutions, the OOA is split into factors dominated by \( m/z \) 43 and \( m/z \) 44.

Other non-physical mass spectra appear in solutions with four or more factors. For example, a spectrum dominated by \( m/z \) 15, also observed in the 3-vector solutions, is common in the vector-matrix solutions. In addition, a new, nonphysical factor not observed in the 3-vector solutions is dominated by \( m/z \)'s 67, 81, and 95 (Fig. H.12). This series of \( m/z \)'s is present in mass spectra of dienes, alkynes, and cycloalkenes (McLaugherty and Turecek, 1993), but the spectra of real compounds contain more than just these three fragments. Thus, these spectra are not known to represent a real aerosol type. No new, meaningful factors are identified in solutions with five or more factors. Consequently, we reject all of the unconstrained vector-matrix solutions.
H.5 Constraining the vector-matrix model using multiple linear regression

Multiple linear regression solves the same model as Eq. (H.1), but the factor compositions (i.e., the $c_j$ values for $p$ factors) must be provided a priori. In this work, the $p = 4$ factor compositions came from the four HR-MS mass spectra. The size-resolved mass spectrum measured at each time step was fit separately, i.e., $36 \times 1366$ fits are performed. Regressions were performed in Igor Pro v. 6.21 (Wavemetrics, Portland, OR) using the Levenberg-Marquard least-squares method. In addition, the fit coefficients ($a_j$) were constrained to be non-negative, matching the constraints of the solutions of the vector-matrix model solved by ME-2.

H.5.1 Solutions of the multiple linear regression

The multiple linear regression of the HR-MS factors to the size-resolved composition data failed to fit most of the mass spectra. The regression was only successful (i.e., the matrix was not singular and the fit converged in the allowed number of steps) for 22.5% of the mass spectra. Most of these spectra are from particle size bins with diameters between 50 and 700 nm that have moderate mass spectral signal (0.5–2 µg/m$^3$/decade $d_{va}$). In contrast, the fit did not converge for the majority of the spectra (77%). The non-convergent fits include almost all of the data for particle size bins with $d_{va} < 50$ nm and $d_{va} > 700$ nm, and also particle size bins with $d_{va}$ between 50 and 700 nm that have low signal (< 0.5 µg/m$^3$/log nm). Finally, the remaining 0.5% of mass spectral fits failed because the fitting matrix was singular (i.e., the mass spectra are linearly dependent, so an infinite number of solutions, are possible). The matrix singularity occurs during high-BBOA events, most likely because of near-colinearity between the BBOA and HOA spectra. Unfortunately, these failures occur when we are the most interested in the composition of the aerosol and whether the size distribution is evolving. In total, fitting the size-
resolved mass spectra with the HR-MS mass spectra by multiple linear regression fails for 77.5% of the points. Furthermore, when the regression fits fail, no result is given for these points. Thus, this method is unable to provide factorization results from important, high-mass-loading events and cannot be used for this dataset.
**Fig. H.1.** Particle time-of-flight (PToF) data at (a) $m/z$ 43 and (b) $m/z$ 16. The raw data (dashed line) have a high background, estimated from the shaded region. The background estimation regions are selected for times before particles are expected to arrive and after no particles are expected to arrive at the vaporizer. Subtracting the average background level gives the solid line. When gas-phase signal is expected before particle arrival times, as for O$^+$ at $m/z$ 16 in (b), only the later background region is used for background subtraction.
Fig. H.2. Fraction of points in the 3D data matrix whose error was increased to a minimum value of 1 ion and the average increase of those values for each (a) m/z and (b) size bin. Average increases for m/z’s were highest for a few m/z’s > 85 (black lines). Errors were increased more frequently (grey bars) for particle size bins with the largest and smallest diameters.
Fig. H.3. Average signal-to-noise ratio (SNR) of each m/z (left axis) at each particle size (bottom axis). Size-m/z combinations are shaded by SNR. Combinations with SNR < 1.5 are considered “weak” and are shaded in grey. Combinations with SNR ≥ 1.5 are considered “strong” and are shaded in color. White areas denote m/z’s that are not assigned organic signal.
**Fig. H.4.** Mass size distribution \( (dM/d\log d_{va}) \) for the best solution of the constrained vector-matrix model plotted vs. \( d_{va} \) on a log scale on the y-axis and vs. sampling date on the x-axis. Grey pixels have zero signal.
Fig. H.5. Case study event on 24 March 2006. (a) Mass size distributions (dM/d\log d_{va}) of the 4 factors in the best solution of the vector-matrix model, normalized so that each size distribution has unit area, plotted vs. \(d_{va}\) on a log scale on the y-axis and vs. sampling time on the x-axis. The data have been binomially smoothed by one point each in time and size. (b) Mass size distributions (dM/d\log d_{va}) that have not been normalized, plotted vs. \(d_{va}\) on a log scale on the y-axis and vs. sampling time on the x-axis. In both panels, light-grey pixels have zero signal.
Fig. H.6. \( Q/Q_{exp} \) values for 50 seed solutions of 4 factors (a–c) and 5 factors (d–f) grouped into families with similar factors. Unweighted \( Q/Q_{exp} \) values are in the top row of each panel (blue markers), and \( Q/Q_{exp} \) values as computed from the factorization are shown in the bottom row of each panel (red markers). Solutions were calculated for (a, d) the 3-vector model solved with PMF3, (b, e) the 3-vector model solved with ME-2, and (c, f) the unconstrained vector-matrix model solved with ME-2.
Fig. H.7. Solution of the 3-vector model with 4 factors and the lowest $Q/Q_{exp}$ values. Two of the factors have mass spectra dominated by $m/z$'s 44 and 43, and two of the factors represent HOA and biomass-burning organic aerosol BBOA. (a) Mass spectrum of each factor plotted vs. ion mass-to-charge ratio ($m/z$). Mass spectra are normalized to sum to 1. (b) Mass size distribution ($dM/d\log(d_{va})$) plotted vs. particle vacuum-aerodynamic diameter ($d_{va}$) on a log scale. Size distributions are normalized so that the area under each curve sums to 1. (c) Mass contribution of each factor plotted vs. sampling date. The scale for LOA has been expanded to show the structure during low-concentration periods.
Fig. H.8. Solution of the 3-vector model with 4 factors that include two with similar mass spectra. The four factors are OOA, HOA, and two BBOA factors. (a) Mass spectrum of each factor plotted vs. ion mass-to-charge ratio (m/z). Mass spectra are normalized to sum to 1. (b) Mass size distribution (dM/dlog\(d_{va}\)) plotted vs. particle vacuum-aerodynamic diameter (\(d_{va}\)) on a log scale. Size distributions are normalized so that the area under each curve sums to 1. (c) Mass contribution of each factor plotted vs. sampling date. The scale for LOA has been expanded to show the structure during low-concentration periods.
Fig. H.9. Five-factor solution of the constrained vector-matrix model. Four a priori mass spectra were provided as starting guesses: OOA, HOA, and BBOA from the HR-MS solution, and LOA from the best solution of the 3-vector model (Fig. 3). The a priori spectra were not allowed to vary ($\beta = 0$). The fifth factor has a mass spectrum that is HOA-like, but with the signal shifted to $m/z$'s $\geq 43$. (a) Mass spectrum of each factor plotted vs. $m/z$. Mass spectra are normalized to sum to 1. (b) Mass size distribution $(dM/d\log d_{uv})$ plotted vs. $d_{uv}$ on a log scale on the y-axis and vs. sampling date on the x-axis. Grey pixels have zero signal.
**Fig. H.10.** Schematic representation of 2D factorizations of datasets that include size distribution information (a–d) or size-resolved aerosol composition (e–f). Studies that included these factorizations are summarized in Table 1. Details are given in Table S2.
**Fig. H.11.** Time to calculate each solution of the 3-vector and vector-matrix models of our 3.5 x $10^6$ element matrix with 1 to 8 factors using the PMF3 or ME-2 algorithm. Each mark show the time to calculate a solution from a starting seed. The calculation of each solution begins with loading the data and error matrices; this takes ~ 1 minute. Details about the computer used for these calculations and the values of some ME-2 variables are shown in Table H.9. Values are meant only as a rough guide because many elements of the script file can impact the speed of calculation, and we generally did not try to optimize the running speed.
Fig. H.12. Four-factor solution of the unconstrained vector-matrix model, including a factor with a mass spectrum dominated by m/z’s 67, 81, and 95. (a) Mass spectrum of each factor plotted vs. m/z. Mass spectra are normalized to sum to 1. (b) Mass size distribution (dM/dlogdva) plotted vs. dva on a log scale on the y-axis and vs. sampling date on the x-axis. Grey pixels have zero signal.
Table H.1. Downweighting multipliers for $m/z$’s with interference from late-arriving air molecules at $m/z$ 29 (N$^{15}$N) and $m/z$ 44 (CO$_2$). The downweighting magnitude is 100 at the smallest nominal particle sizes and then decreases linearly over four size points to a magnitude of 2. The larger population of N$^{15}$N molecules has a higher probability of causing a substantial signal at later nominal particle sizes, and so stronger downweighting is applied to larger nominal sizes for $m/z$ 29 than $m/z$ 44.

<table>
<thead>
<tr>
<th>Nominal Particle Size ($d_{va}$, nm)</th>
<th>Downweighting Multiplier for $m/z$ 29</th>
<th>Downweighting Multiplier for $m/z$ 44</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>19</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>26</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>34</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>44</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>55</td>
<td>100</td>
<td>75.5</td>
</tr>
<tr>
<td>68</td>
<td>100</td>
<td>51</td>
</tr>
<tr>
<td>82</td>
<td>100</td>
<td>26.5</td>
</tr>
<tr>
<td>98</td>
<td>100</td>
<td>2</td>
</tr>
<tr>
<td>115</td>
<td>75.5</td>
<td>2</td>
</tr>
<tr>
<td>135</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>26.5</td>
<td>2</td>
</tr>
<tr>
<td>179</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
**Table H.2.** Summary of research that applied 2-dimensional (2D) factorization techniques to aerosol size distributions (first four groups) or size-resolved chemical composition of aerosols (last two groups). Details are given in Tables S3–8, and schematic representations of the matrix factorization are shown in Fig. S10.

<table>
<thead>
<tr>
<th>Matrix-Row Representation</th>
<th>Matrix-Row Contents</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>Total size distribution (number, surface area, or mass)</td>
<td>(Kim et al., 2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Zhou et al., 2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Zhou et al., 2005b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Ogulei et al., 2006a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Ogulei et al., 2007b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Chan and Mozurkewich, 2007a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Yue et al., 2008)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Beddows et al., 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Costabile et al., 2009)</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>Total size distribution, Gas-phase speciation</td>
<td>(Wahlin et al., 2001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Thimmaiah et al., 2009)</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>Total size distribution, Chemical speciation of the bulk aerosol</td>
<td>(Ruuskanen et al., 2001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Larson et al., 2006)</td>
</tr>
<tr>
<td><img src="image4.png" alt="Image" /></td>
<td>Total size distribution, Chemical speciation of the bulk aerosol, Gas-phase speciation</td>
<td>(Zhou et al., 2005a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Ogulei et al., 2006b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Ogulei et al., 2007a)</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td>Chemical speciation of size-resolved aerosol, each size factored separately</td>
<td>(Yakovleva et al., 1999)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Dillner et al., 2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Han et al., 2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Yatkin and Bayram, 2008)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Gietl and Klemm, 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Karanasiou et al., 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Kleeman et al., 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Srivastava et al., 2009)</td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td>Chemical speciation of size-resolved aerosol, all sizes factored together</td>
<td>(Amato et al., 2009)</td>
</tr>
</tbody>
</table>

*In these studies, the data were also arranged as a 3-dimensional (3D) matrix and factored using a 3D model. Details of these datasets and models are presented in Table 1.*
<table>
<thead>
<tr>
<th>Sampling Location and Year</th>
<th>Citation</th>
<th>Instrumentation</th>
<th>Chemical Speciation</th>
<th>Sizes (μm)</th>
<th>Sampling Time</th>
<th>Factorization Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seattle, WA 2000-2001</td>
<td>(Kim et al., 2004)</td>
<td>DMPS</td>
<td></td>
<td>0.02–0.4</td>
<td>1 hr</td>
<td>PMF, UNMIX</td>
</tr>
<tr>
<td>Pittsburgh, PA 2001</td>
<td>(Zhou et al., 2004)</td>
<td>SMPS, APS</td>
<td></td>
<td>0.003–2.5</td>
<td>15 min</td>
<td>PMF</td>
</tr>
<tr>
<td>Pittsburgh, PA 2001-2002</td>
<td>(Zhou et al., 2005b)</td>
<td>SMPS, APS</td>
<td></td>
<td>0.003–2.5</td>
<td>15 min</td>
<td>PMF</td>
</tr>
<tr>
<td>Reston, VA (indoors) 1999-2000</td>
<td>(Ogulei et al., 2006a)</td>
<td>SMPS, APS</td>
<td></td>
<td>0.01–20</td>
<td>30 min</td>
<td>PMF</td>
</tr>
<tr>
<td>Egbert, ON, Vancouver, BC 1999-2003</td>
<td>(Chan and Mozurkewich, 2007a)</td>
<td>SMPS</td>
<td></td>
<td>0.006–0.3</td>
<td>5 min</td>
<td>APCA</td>
</tr>
<tr>
<td>Hamilton, ON Simcoe, ON 1999-2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buffalo, NY 2004</td>
<td>(Ogulei et al., 2007b)</td>
<td>EEPS</td>
<td></td>
<td>0.006–0.3</td>
<td>1-5 sec</td>
<td>PMF</td>
</tr>
<tr>
<td>Erfurt, Germany 1997-2001</td>
<td>(Yue et al., 2008)</td>
<td>DMPS, OLAS</td>
<td></td>
<td>0.01 – 3.0</td>
<td>1 hr</td>
<td>PMF</td>
</tr>
<tr>
<td>London, UK 2005</td>
<td>(Beddows et al., 2009)</td>
<td>SMPS</td>
<td></td>
<td>0.012–0.437, 0.015–0.661</td>
<td>1 hr, 6 hr</td>
<td>Cluster Analysis</td>
</tr>
<tr>
<td>Leipzig, Germany 2005-2006</td>
<td>(Costabile et al., 2009)</td>
<td>TDMPS, SMPS</td>
<td></td>
<td>0.003–0.9</td>
<td>30 min</td>
<td>PCA</td>
</tr>
</tbody>
</table>
b) Acronym Sampling Technique

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Sampling Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS</td>
<td>Aerodynamic Particle Sizer</td>
</tr>
<tr>
<td>DMPS</td>
<td>Differential Mobility Particle Sizer</td>
</tr>
<tr>
<td>EEPS</td>
<td>Engine Exhaust Particle Spectrometer</td>
</tr>
<tr>
<td>OLAS</td>
<td>Optical Laser Aerosol Spectrometer</td>
</tr>
<tr>
<td>SMPS</td>
<td>Scanning Mobility Particle Sizer</td>
</tr>
<tr>
<td>TDMPS</td>
<td>Twin Differential Mobility Particle Sizer</td>
</tr>
</tbody>
</table>

c) Acronym Factorization Technique

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Factorization Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>APCA</td>
<td>Absolute Principal Component Analysis</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PMF</td>
<td>Positive Matrix Factorization</td>
</tr>
<tr>
<td>UNMIX</td>
<td>Unmix multivariate receptor model</td>
</tr>
</tbody>
</table>

Table H.3. Details of research that reports application of 2D factorization techniques to aerosol size distributions. Table H.3b expands the acronyms of instrumental techniques shown in Table H.3a. Table H.3c expands the acronyms of factorization techniques shown in Table H.3a.
<table>
<thead>
<tr>
<th>Sampling Location and Year</th>
<th>Citation</th>
<th>Instrumentation</th>
<th>Chemical Speciation</th>
<th>Sizes (μm)</th>
<th>Sampling Time</th>
<th>Factorization Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copenhagen and Odense, Denmark 1999</td>
<td>(Wahlin et al., 2001)</td>
<td>DMPS, CPC</td>
<td>CO, NO&lt;sub&gt;x&lt;/sub&gt;</td>
<td>0.006–0.7</td>
<td>30 min</td>
<td>COPREM</td>
</tr>
<tr>
<td>Prague, Czech Republic 2008</td>
<td>(Thimmaiah et al., 2009)</td>
<td>SMPS</td>
<td>CO, NO&lt;sub&gt;x&lt;/sub&gt;, SO&lt;sub&gt;2&lt;/sub&gt;, O&lt;sub&gt;3&lt;/sub&gt;, CH&lt;sub&gt;4&lt;/sub&gt;</td>
<td>0.018–0.723</td>
<td>1 hr</td>
<td>PMF</td>
</tr>
</tbody>
</table>

**Table H.4.** Details of research that reports application of 2D factorization techniques to datasets pairing aerosol size distributions and particle-phase composition. Table H.2b expands the acronyms of instrumental techniques shown in Table H.4a. Table H.4c expands the acronyms of factorization techniques shown in Table H.4a.
<table>
<thead>
<tr>
<th>Sampling Location and Year</th>
<th>Citation</th>
<th>Instrumentation</th>
<th>Chemical Speciation</th>
<th>Sizes (μm)</th>
<th>Sampling Time</th>
<th>Factorization Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akjnaar, Netherlands Erfurt, Germany Helsinki, Finland 1996-1997</td>
<td>(Ruuskanen et al., 2001)</td>
<td>DMPS SMPS OLAS EAS CPC Harvard Impactor</td>
<td>PM$_{2.5}$ mass</td>
<td>0.01–10</td>
<td>1 hr</td>
<td>PCA</td>
</tr>
<tr>
<td>Seattle, WA 2000-2003</td>
<td>(Larson et al., 2006)</td>
<td>DMPS APS STN</td>
<td>PM$_{2.5}$ mass</td>
<td>0.02–5</td>
<td>24 hr</td>
<td>PMF with additional constraints, solved with ME-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NH$_4^+$, NO$_3^-$, SO$_4^{2-}$, K$^+$, Na$^+$</td>
<td>17 elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EC, OC</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) Acronym Sampling Technique

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Sampling Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS</td>
<td>Aerodynamic Particle Sizer</td>
</tr>
<tr>
<td>CPC</td>
<td>Condensation Particle Counter</td>
</tr>
<tr>
<td>DMPS</td>
<td>Differential Mobility Particle Sizer</td>
</tr>
<tr>
<td>EAS</td>
<td>Electrical Aerosol Spectrometer</td>
</tr>
<tr>
<td>OLAS</td>
<td>Optical Laser Aerosol Spectrometer</td>
</tr>
<tr>
<td>SMPS</td>
<td>Scanning Mobility Particle Sizer</td>
</tr>
<tr>
<td>STN</td>
<td>Speciation Trends Network</td>
</tr>
</tbody>
</table>

c) Acronym Factorization Technique

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Factorization Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PMF</td>
<td>Positive Matrix Factorization</td>
</tr>
<tr>
<td>ME-2</td>
<td>Multilinear Engine 2</td>
</tr>
</tbody>
</table>

**Table S5.** Details of research that reports application of 2D factorization techniques to datasets pairing aerosol size distributions and gas-phase composition. Table S3b expands the acronyms of instrumental techniques shown in Table S5a. Table S5c expands the acronyms of factorization techniques shown in Table S5a.
<table>
<thead>
<tr>
<th>Sampling Location and Year</th>
<th>Citation</th>
<th>Instrumentation</th>
<th>Chemical Speciation</th>
<th>Sizes (μm)</th>
<th>Sampling Time</th>
<th>Factorization Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pittsburgh, PA 2001</td>
<td>(Zhou et al., 2005a)</td>
<td>SMPS</td>
<td>SO$_2^-$, NO$_3^-$</td>
<td>0.003–2.5</td>
<td>15 min</td>
<td>PLS, PMF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>APS</td>
<td>11 elements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O$_3$, NO, NO$_x$,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SO$_2$, CO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baltimore, MD 2002</td>
<td>(Ogulei et al., 2006b)</td>
<td>SMPS</td>
<td>PM$_{2.5}$ mass</td>
<td>0.00965–2.458</td>
<td>1 hr</td>
<td>PLS, PMF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>APS</td>
<td>NO$_3^-$, SO$_4^{2-}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SEAS</td>
<td>EC, OC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11 elements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CO, NO, NO$_x$,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SO$_2$, O$_3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rochester, NY 2004-2005</td>
<td>(Ogulei et al., 2007a)</td>
<td>SMPS</td>
<td>PM$_{2.5}$ mass</td>
<td>0.012–0.470</td>
<td>1 hr</td>
<td>PMF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CO, O$_3$, SO$_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) Acronym Sampling Technique

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Sampling Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS</td>
<td>Aerodynamic Particle Sizer</td>
</tr>
<tr>
<td>SMPS</td>
<td>Scanning Mobility Particle Sizer</td>
</tr>
<tr>
<td>SEAS</td>
<td>Semi-continuous Elements in Aerosol</td>
</tr>
</tbody>
</table>

c) Acronym Factorization Technique

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Factorization Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>Partial Least Squares</td>
</tr>
<tr>
<td>PMF</td>
<td>Positive Matrix Factorization</td>
</tr>
</tbody>
</table>

**Table H.6.** Details of research that reports application of 2D factorization techniques to datasets that combine aerosol size distributions with gas- and particle-phase composition. Table H.6b expands the acronyms of instrumental techniques shown in Table H.6a. Table H.6c expands the acronyms of factorization techniques shown in Table H.6a.
<table>
<thead>
<tr>
<th>Sampling Location and Year</th>
<th>Citation</th>
<th>Instrumentation</th>
<th>Chemical Speciation</th>
<th>Sizes (μm)</th>
<th>Sampling Time</th>
<th>Factorization Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riverside, CA 1991</td>
<td>(Yakovleva et al., 1999)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Stationary indoor monitors, Stationary ambient monitors, Personal exposure monitors</td>
<td>18 elements</td>
<td>PM&lt;sub&gt;2.5&lt;/sub&gt;, PM&lt;sub&gt;10&lt;/sub&gt;, personal PM&lt;sub&gt;10&lt;/sub&gt; (2 sizes, 5 types)</td>
<td>12 hr</td>
<td>PMF</td>
</tr>
<tr>
<td>Houston, TX 2000</td>
<td>(Dillner et al., 2005)</td>
<td>MOUDI</td>
<td>SO&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;2-&lt;/sup&gt;, NO&lt;sub&gt;3&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;, Cl&lt;sup&gt;-&lt;/sup&gt;, NH&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;+&lt;/sup&gt;, EC, OM</td>
<td>0.056–1.8 (6 ranges)</td>
<td>24 hr</td>
<td>Cluster Analysis</td>
</tr>
<tr>
<td>Gosan, Korea 2002</td>
<td>(Han et al., 2006)</td>
<td>DRUM</td>
<td>19 elements</td>
<td>0.07–12 (8 ranges)</td>
<td>3 hr</td>
<td>PMF</td>
</tr>
<tr>
<td>Izmir, Turkey 2004-2005</td>
<td>(Yatkin and Bayram, 2008)</td>
<td>Dichotomous sampler</td>
<td>16 elements</td>
<td>PM&lt;sub&gt;2.5&lt;/sub&gt;, PM&lt;sub&gt;10&lt;/sub&gt; (2 sizes)</td>
<td>24 hr</td>
<td>PMF, CMB</td>
</tr>
<tr>
<td>Münster, Germany 2006-2007</td>
<td>(Gietl and Klemm, 2009)</td>
<td>Berner impactor</td>
<td>SO&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;2-&lt;/sup&gt;, NO&lt;sub&gt;3&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;, Cl&lt;sup&gt;-&lt;/sup&gt;, NH&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;+&lt;/sup&gt;, Na&lt;sup&gt;+&lt;/sup&gt;, Ca&lt;sup&gt;2+&lt;/sup&gt;, Mg&lt;sup&gt;2+&lt;/sup&gt;, EC, OC</td>
<td>0.053–10 (5 ranges)</td>
<td>5-7.5 hr</td>
<td>PMF</td>
</tr>
<tr>
<td>Athens, Greece 2002</td>
<td>(Karanasiou et al., 2009)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Custom impactors aethalometer</td>
<td>SO&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;2-&lt;/sup&gt;, BC</td>
<td>PM&lt;sub&gt;2&lt;/sub&gt;, PM&lt;sub&gt;10-2&lt;/sub&gt;</td>
<td>24 hr</td>
<td>PMF</td>
</tr>
</tbody>
</table>

<sup>a</sup> Indicates a review or summary paper.
### Table H.7. Details of research that reports application of 2D factorization techniques to size-resolved aerosol composition data.

In these studies, the aerosol composition from each size range is factored separately. Table H.7b expands the acronyms of instrumental techniques shown in Table H.7a. Table H.7c expands the acronyms of factorization techniques shown in Table H.7a.

*a* In these works, the data was also arranged as a 3D matrix and factored using a 3D model; details of these datasets and models are presented in Table 3.1.

<table>
<thead>
<tr>
<th>Sampling Location and Year</th>
<th>Citation</th>
<th>Instrumentation</th>
<th>Chemical Speciation</th>
<th>Sizes (μm)</th>
<th>Sampling Time</th>
<th>Factorization Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sacramento, Modesto, and Bakersfield, CA 2000-2001</td>
<td>(Kleeman et al., 2009)</td>
<td>MOUDI</td>
<td>8 molecular organic tracers</td>
<td>0.055–1.8 (6 ranges)</td>
<td>8, 12 hr</td>
<td>Custom source apportionment algorithm</td>
</tr>
<tr>
<td>Delhi, India 2005-2006</td>
<td>(Srivastava et al., 2009)</td>
<td>Cascade Impactor</td>
<td>11 metals</td>
<td>0.7–10.9 (2 ranges)</td>
<td>24 hr</td>
<td>CMB</td>
</tr>
</tbody>
</table>

#### b) Acronym Sampling Technique
- DRUM: Davis Rotating Unit for Monitoring
- MOUDI: Micro-Orifice, Uniform Deposit Impactor

#### c) Acronym Factorization Technique
- CMB: Chemical Mass Balance
- PCA: Principal Component Analysis
- PMF: Positive Matrix Factorization
<table>
<thead>
<tr>
<th>Sampling Location and Year</th>
<th>Citation</th>
<th>Instrumentation</th>
<th>Chemical Speciation</th>
<th>Sizes (µm)</th>
<th>Sampling Time</th>
<th>Factorization Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcelona, Spain 2003-2007</td>
<td>(Amato et al., 2009)</td>
<td>High-volume samplers</td>
<td>NO$_3^-$, Cl$^-$, NH$_4^+$</td>
<td>PM$<em>{10}$, PM$</em>{2.5}$, PM$_1$</td>
<td>24 hr</td>
<td>PMF with pulling equations in ME-2</td>
</tr>
</tbody>
</table>

b)  

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Factorization Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME-2</td>
<td>Multilinear Engine 2</td>
</tr>
<tr>
<td>PMF</td>
<td>Positive Matrix Factorization</td>
</tr>
</tbody>
</table>

**Table H.8.** Details of research that reports application of 2D factorization techniques to size-resolved aerosol composition data. In these studies, the aerosol composition from all size ranges is combined in one matrix and factored simultaneously. Table H.8b expands the acronyms of factorization techniques shown in Table H.8a.
Table H.9. Configuration of the computer and ME-2 algorithm used to solve the factorizations. No attempt was made to adjust the settings in the ME-2 control file to increase the speed of the calculations.

<table>
<thead>
<tr>
<th>Computer Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM</td>
</tr>
<tr>
<td>Processor</td>
</tr>
<tr>
<td>Operating System</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ME-2 Control File Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convergence test level 1</td>
</tr>
<tr>
<td>Convergence test level 2</td>
</tr>
<tr>
<td>Convergence test level 3</td>
</tr>
<tr>
<td>cgresets</td>
</tr>
<tr>
<td>Precondition mode</td>
</tr>
</tbody>
</table>