Visual Steering and Verification of Mass Spectrometry Data Factorization in Air Quality Research

Daniel Engel, *Member, IEEE*, Klaus Greff, Christoph Garth, *Member, IEEE*, Keith Bein,
Anthony Wexler, Bernd Hamann, *Member, IEEE* and Hans Hagen, *Member, IEEE*

Abstract— The study of aerosol composition for air quality research involves the analysis of high-dimensional single particle mass spectrometry data. We describe, apply, and evaluate a novel interactive visual framework for dimensionality reduction of such data. Our framework is based on non-negative matrix factorization with specifically defined regularization terms that aid in resolving mass spectrum ambiguity. Thereby, visualization assumes a key role in providing insight into and allowing to actively control a heretofore elusive data processing step, and thus enabling rapid analysis meaningful to domain scientists. In extending existing black box schemes, we explore design choices for visualizing, interacting with, and steering the factorization process to produce physically meaningful results. A domain-expert evaluation of our system performed by the air quality research experts involved in this effort has shown that our method and prototype admits the finding of unambiguous and physically correct lower-dimensional basis transformations of mass spectrometry data at significantly increased speed and a higher degree of ease.

Index Terms—Dimension reduction, mass spectrometry data, validation and verification of matrix factorization, visual encodings of numerical error metrics, multi-dimensional data visualization.

1 Introduction

Atmospheric particles increase morbidity and mortality in polluted urban areas and alter the Earth's radiative energy balance related to climate change. Innovative instruments are now capable of chemically analyzing individual airborne particles in real time, providing unprecedented rich data sets for air quality research. Typical instruments analyze hundreds of thousands of particles over a few weeks to year-long period, with each measurement reporting typically 250 ion mass/charge ratios in the mass spectrum of each particle. Current analysis techniques employ clustering the spectra and averaging of similar spectra to represent common composition. However, these clustering algorithms have a number of shortcomings. For instance, averaging obscures important sources or atmospheric processes that the particles have undergone. Moreover, the distance measures used by these algorithms are known to misclassify spectra due to isobaric interferences. *Isobaric interference* refers to the fact that different physical sources contribute to the same dimension in the mass spectrum, rendering its values ambiguous as the contributing components and their magnitudes are unknown. However, since different compounds in the gas and particle phase manifest themselves in multiple ways in the particle mass spectra and in the gas phase composition, the data are often composed of independent or nearly independent components that fundamentally characterize the particle and moreover the air parcel by their combination. As a result, understanding these vast high-dimensional data sets can be facilitated by interpreting the data based on their independent compounds. Consequently, a basis transformation is needed that resolves ambiguity by expressing the data into coefficients of its latent physical components.

The problem presented here is known as that of blind source separation [7]. Given data that is derived from a combination of unknown sources in unknown occurrences, the goal is to factor out both unknowns given only an estimate of the number of sources and an assumption of the data model that defines how sources combine. In particular, sources are not mutually orthogonal which leads to ambiguity in the data. Thereby, the sources and their occurrence represent the actual "hidden" information that is to be factored out from the given mixture. In the context of atmospheric processes, the unknown sources represent the latent components that appear independently in aerosols. These form the actual, unambiguous, and lower-dimensional data basis. Based on the physical model of particle ionization, each measured aerosol mass spectrum can be described by the linear combination of the mass spectra of its components. As a consequence, the latent independent variable basis of the data and the corresponding coefficients that derive the data are, in theory, exactly the independent physical components and their occurrences in the aerosols. However, in practice, extracting these basis vectors and coefficients so that their physical composition can be interpreted proves difficult. In particular, methods like independent components analysis [6] do not account for non-negativity in the mass spectrum and physical source mixture model. As a result, air quality researchers are ambivalent about using dimension reduction when the resulting data basis does not show a relationship to air pollutant emissions and their atmospheric processing.

The authors, from the domains of air quality research and visualization, jointly studied this problem to investigate the positive influence of visualization on solving this problem. Dealing effectively with these vast, high-dimensional data sets necessitates the practical need for visualizing the process of dimension reduction, thereby instigating a research direction for visualization in which little prior work exists. We show in this paper that visualization is essential in providing air quality researchers with the means of finding unambiguous physically correct lower-dimensional basis transformations of their data. Our method involves the visualization of non-convex, multi-criteria, and non-negative matrix factorization. Further, our method entails a visual interface to this optimization process that (i) allows the atmospheric scientist to be "in the loop" of the computation, (ii) provides direct visual feedback of the optimization process, and (iii) enables controlled refinement of its solution. We introduce domain-specific visual encodings and interactive mechanisms of matrix factorization that provide the means to incorporate expert knowledge into numerical optimization and to steer this process toward physical meaning while maintaining mathematical rigor. Thereby, we contribute both to the field of air quality research by providing novel means for the research of aerosol source contributions, as well as to the visualization community by laying the groundwork for further research toward enabling physically meaningful and interpretable dimension reduction.

- *D. Engel, K. Greff, C. Garth, H. Hagen are with the University of Kaiserslautern, Germany, E-mail: {d_engel, greff, garth, hagen}@cs.uni-kl.de.*
- *B. Hamann is with the Institute for Data Analysis and Visualization (IDAV), Department of Computer Science, University of California, Davis, CA, USA, E-mail: hamann@cs.ucdavis.edu.*
- *K. Bein and A. Wexler are with the Air Quality Research Center (AQRC), University of California, Davis, CA, USA, E-mail: {kjbein, aswexler}@ucdavis.edu.*

The remainder of the paper is structured as follows. Section 2 discusses related work in dimension reduction, visualization, and air quality research, while Section 3 provides the necessary application background, task description, and requirements for our effort. Our framework, consisting of data factorization, domain-specific visual encodings and interaction mechanisms, is described in Section 4. In Section 5, this framework is applied to the factorization of biomass combustion particle spectra and evaluated with respect to its ability to produce new insights to the application of air quality research. Finally, concluding remarks are given in Section 6.

2 Related work

Work related to the scope of this paper can be found in the fields of dimension reduction, visualization, and air quality research; we provide a brief overview in the following.

Dimension reduction

Methods for dimension reduction compute a mapping from high- to low-dimensional space. Thereby, data is decomposed into a set of new coordinates, acting as coefficients to a different basis that is more suitable with respect to data properties. In visualization, dimension reduction is commonly applied as a means of finding a lower-dimensional data embedding that best reflects distance relationships between high-dimensional points. Here, the focus lies not on the properties of the basis but on their coefficients that act as an abstraction to the high-dimensional data. In contrast, we are concerned with the basis of this mapping, as its interpretation is essential in air quality research.

Methods like multi-dimensional scaling [36] or manifold learning [35, 32] define data bases in inner product space. These bases prove hard to interpret as their dimensionality equals the number of data points. While principal components analysis [29] finds bases in data space, it is restricted to orthogonal basis vectors. Overcoming this restriction, independent components analysis [6, 14] finds non-orthogonal independent data bases. However, methods based on higher-order statistics are generally unsuited for our problem domain, as single particle mass spectrometry (SPMS) data is non-negative by nature. The methods mentioned above make use of spectral decompositions and are thereby incapable of factorizing non-negative data into a non-negative basis and non-negative coefficients [5].

In comparison to classical statistical approaches, matrix factorization (MF) methods offer more degrees of freedom for defining optimization goals. In particular, non-negative matrix factorization (NMF) has recently received great attention because it is capable of computing non-negative basis transformations. Non-negativity is an integral property for application areas that investigate physical phenomena described by non-negative measurements or mixtures, as it is the case for the type of air quality data we are concerned with. In this paper, we make use of the works of [20] and [39]. The former provides a framework for alternating non-negative least squares, while the latter shows how the use of a decorrelation regularization term derives independent components in non-negative data. In contrast to previous work, no matrix inversion is necessary for this computation. Other work offers a convex model to NMF but is constraint in expressing its basis as a convex combination of data points [9].

As MF methods are based on numerical optimization, their drawbacks lie with computation and convergence speed, as well as their proneness to "get stuck" in local minima in the case of non-convex optimization problems. Finally, all dimension reduction methods share the common problem of finding a basis transformation that is not only numerically correct but also physically interpretable by the domain expert. With complex physical restrictions that are not (yet) well-defined, it is impossible to extract a physically correct data basis through numerical optimization alone. After all, in air quality research, as well as in most scientific applications, the motivation for exploratory data analysis is that data properties are part of open research questions. Consequently, the scientist has to be involved in interacting with computations; this insight motivates the present work.

Visualization

While visual steering of exploration [34] and simulations [21, 22, 38] have become a well-established research areas in the field of visualization, the visual steering of practical engineering optimization has not been the focus of previous studies. However, this need is clearly documented [25] and recent advances from the engineering community give empirical evidence to the benefits of interactive visualization-based strategies to support engineering design optimization and decision-making [4].

Recently, novel techniques have been introduced to the visualization community that enable user interaction in dimension reduction and have demonstrated great success in application areas. For example, piecewise Laplacian-based projection (PLP) [28] allows the user to interact directly with the mapping process by providing means to adjust neighborhood and distance approximations. Thus, the user can implicitly redefine the basis for dimension reduction. However, the application of PLP does not involve analysis of this basis or the mapping error. Previous methods have used dimension reduction as a means to visualize data. In contrast, our application requires visualization as a means to steer dimension reduction toward physical correctness. In this regard, air quality research may prove to trigger a new research domain for visualization.

Techniques to visualize matrix factorization can, in part, be based upon existing research in multi-dimensional data visualization [37, 11]. Driven by applications, research focuses on better representation of specific data properties (e.g., scientific point cloud data [27]), better incorporation of domain-appropriate analysis techniques (like brushing and filtering [17]), or computational speed gains [15]. Opposed to visualizing data relationships by dimension reduction, value visualizations like heat maps, glyphs, scatter plot matrices, and parallel coordinates, are regarded as the dominant approach to multi-dimensional data visualization. Research in this area has focused on enhanced cluster visualization [18, 44, 1], brushing techniques [8, 13], and abstraction [24]. Clutter reduction through dimension ordering [30, 40, 10] is one of the most promising approaches to enable data comprehension.

Visualization in air quality research

The air quality research community has recognized the need for tools to assist the interpretation of data from single particle mass spectrometers that generate many spectra of high dimensionality. Published under the synonym positive matrix factorization, NMF has been used for classification of airborne particle types [19]. However, the focus has not been on deriving an independent basis or visualizing the result. As of now, the dominant approach to mass spectrometry data analysis is to apply clustering methods. These methods are commonly based on hierarchical [26], neural network [33], or density [43] schemes that utilize geometric distance measures. Available software packages for mass spectrometry data ([12], [41]) include a variety of data mining methods that focus on clustering and the visual analysis thereof. Recently, a system allowing visual analysis and steering of data clustering has been introduced [42]. Here, better clustering results have been obtained through the incorporation of expert knowledge into data clustering and means for refinement of prior solutions.

3 Requirement analysis

3.1 Application background

Single particle mass spectrometry (SPMS) is used in air quality research to categorize, collect, and analyze aerosols at sampling sites. The research objective is to qualitatively assess air pollution by linking sources to their emitted aerosols, as well as to quantify their toxicity. This allows scientists to assess reliably the degree of pollution that has been emitted by identified entities (such as cars, factories, etc.), as well as their abundance in the atmosphere at specific time and location. One arising scientific problem is that collected samples need to be categorized by emission source (particle type) in order to be accumulated for toxicity analysis. Analyzing the particle's composition is therefore an essential step in this research.

SPMS instruments collect, filter, and characterize aerosols based on their mass spectrum. This spectrum is measured by the instrument based on particle ionization. Aerosols are accelerated through a drift tube and hit by a plasma beam. This causes the particle to ionize and break up into fragments of different compositional levels. The masses of particle fragments are then computed by a time-of-flight analyzer that determines the abundance and mass of each fragment ion. Individual measurements are combined to a mass spectrum for the particle. The mass spectrum represents a function mapping *mass over elemental charge (m/z)* of fragment ions to their abundance in the particle. Discretized in bins of 1 m/z step size, the analyzer captures the first 256 m/z ratios for each aerosol. The resulting histogram data is stored as a 256-dimensional vector, where each coordinate corresponds to the abundance of fragments within the aerosol having an m/z ratio within the dimension's section of the discretized spectrum.

As mass is ambiguous, several fragment ions map to the same m/z dimensions and contribute to each coordinate. This phenomenon is known as *isobaric interference*. Data transformations based on geometric distance metrics are unable to resolve this ambiguity, as they rely on the comparison of coordinates which, in SPMS data, can stem from different physical sources. Consequently, data clustering results are less reliable and the state-of-the-art approach involves verifying each individual (mean) representative spectrum by the scientist. Figure 1 gives an example for this analysis. In the figure, individual peaks are labeled by their source contribution. Ambiguity is resolved by manual analysis and experience.

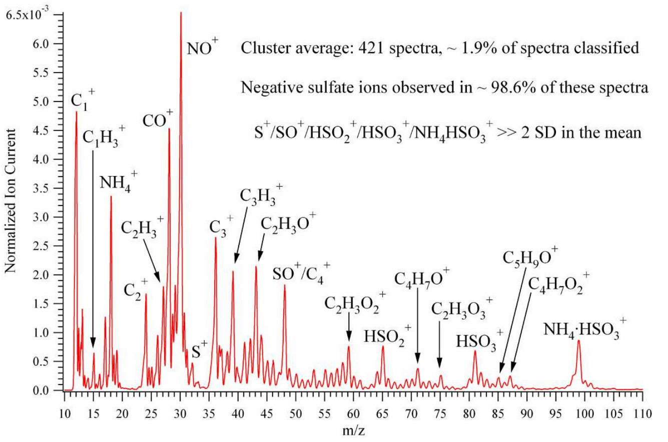

Fig. 1. As of now, atmospheric scientists estimate aerosol composition and the sources that contribute to each m/z dimension based on their experience in investigating mass spectra. Analyzing and classifying thousands of particle spectra can take months even with the help of data clustering.

Ambiguity has created several problems for the application and has in turn triggered basic research problems for visualization. In air quality research, the task does not match the available data. SPMS data is high-dimensional, noisy and ambiguous. The application's goal is categorizing particles by their composition, however, their composition is not reflected by the mass spectrum because of its ambiguity. Consequently, a basis transformation is needed that resolves this ambiguity.

According to the underlying physics, it is assumed that each particle can be described by the linear combination of its fragment ions. Consequently, SPMS data $X \in \mathbb{R}^{(n \times m)}$, holding n particle spectra discretized in m dimensions, can be described by the discretized mass spectra of their fragment ions as a basis B to X, such that

$$
\begin{aligned}
X &= CB + N \text{ and} \qquad (1)\\
X_{i,\bullet} &= \sum_{1 \le j \le |B|} C_{i,j} B_{j,\bullet} + N_{i,\bullet}\ .
\end{aligned}
$$

Here, B is the matrix storing (row-wise) basis vectors, $B_{j,\bullet} \in R^m$, $1 \le j \le |B|$, such that X is derived with the coefficient matrix C and the noise N induced by the instrument. Note that $X_{i,\bullet}$ and $N_{i,\bullet}$, $1 \le i \le n$, refer to the rows of the matrices X and N. Further, all coordinates are positive. The problem is ill-posed and there is no unique solution because (i) C, B, and N are unknown and (ii) any change in B can be undone by changing C accordingly. In particular, adding arbitrary basis vectors to an ideal set B^* with appropriate coefficients (adding up to zero) does not change the solution. From the standpoint of numerical optimization, infinitely many bases exist with which the data can be described. Due to the complexity of the problem, the level of uncertainty involved in data collection, partially unknown machine-dependent physical models of particle fragmentation and noise induced by the instrument, previous methods have failed to deliver physically correct independent data bases.

3.2 Requirements and tasks

Based on the discussions with our co-authors and domain experts, we have identified key requirements for implementing a basis transformation of SPMS data as described in the previous section, as well as the tasks that define the usage of such a system.

Many optimization methods are static in nature and often output their result in combination with a single error measure. For air quality research, however, this is not sufficient. An overall mapping error of "2.05", for example, provides close to no insight for analysts. Atmospheric scientists have extensive experience in interpreting SPMS data, identifying different sources, common features, and noise. Under the mathematically ill-posed problem of finding physically correct bases for the un-mixing of SPMS sources, analytical frameworks are well-advised to draw upon the scientists' knowledge in order to generate more reliable results.

As numerical error measures hardly facilitate understanding of data features or the physical correctness of the factorization basis, visualization of the basis transformation is inevitable when the scientist is to asses its correctness. This visualization should permit a detailed understanding of

- the exact mapping error induced by the basis transformation with respect to the data and
- the basis vectors used in the transformation with respect to the data features.

Further, the scientist requires an interface to perform the following tasks.

- **Analyze basis:**
  In a correct factorization, each basis vector equals the mass spectrum of a stereotypical particle fragment ion that is part of the measured aerosols. The basis should therefore be depicted in a way that makes possible easy comparison between different basis vectors and how they relate to the underlying data features, i.e., the different sources in the data set. In this comparison, analysts should be able to identify the relative peak heights and sparsity of the different basis vectors and verify that the basis describes physically meaningful parts of aerosols.
- **Analyze mapping errors:**
  To verify the correctness of the basis transformation, the scientists needs to analyze how well each part of the data is captured by the factorization. In part, this error may stem from noise or outlier measurements. Both overview and detail is required such that the overall fit, as well as the specific error with respect to certain dimensions or data points, can be assessed.
- **Adjust basis:**
  If the basis is found to be physically incorrect, intuitive means for interaction are required to adjust the basis accordingly. Also, means to account for the level of uncertainty in the basis configuration are desirable, since exact peak ratios may vary depending on measurement parameters but certain conditions of the basis configuration are known and can be specified.

The exact coefficients of the basis are not of immediate interest for verifying the factorization but for successive analysis steps. Instead, the focus lies on the error that is induced by the basis transformation, as well as on the physical correctness of the basis. Computations should be based on the physical data model, mathematically well-founded, and convey physically interpretable results. For visual user feedback, visualization is essential in conveying an understanding of the factorization's intermediate results, as well as to offer an effective interface to verify and control computations with all involved parameters.

4 Method

The goal of the method presented in this section is to provide atmospheric scientists with the means of finding unambiguous physically correct lower-dimensional basis transformations of single particle mass spectrometry (SPMS) data. Our method involves the use of non-negative matrix factorization (NMF) in combination with specific regularization terms to find a basis transformation of SPMS data that minimizes ambiguity. Further, our method entails a visual interface to this optimization process that allows the analyst to be "in the loop" of the computation, provides direct visual feedback of the optimization process, and allows controlled refinement of its solution. By introducing domain-specific visual encodings and interaction mechanisms of SPMS factorization, we provide means to incorporate expert knowledge into the numerical optimization in order to steer this process and to verify the physical correctness of the basis transformation.

4.1 Non-negative matrix factorization

Given n data points of dimension m with non-negative coordinates, $X \in \mathbb{R}_+^{(n\times m)}$ and a positive integer $k \in \mathbb{N}, k > 0$, methods that perform non-negative matrix factorization (NMF) find a factorization of X into a basis $B \in \mathbb{R}_+^{(k\times m)}$ and coefficients $C \in \mathbb{R}_+^{(n\times k)}$, such that

$$||X - CB|| \rightarrow min, \tag{2}$$

where all values in C and B are non-negative.

From the class of existing methods for NMF, we use a combination of [39] and [20], together with two regularization terms of independence and diversity that serve the objective of finding an unambiguous basis and reliable solution, respectively.

The objective of *independence* is understood by taking into account the physical process of particle ionization induced by the SPMS instrument. The contribution of fragment ions to the particle's mass spectrum can be modeled in a hierarchical manner, as larger fragments ionize into smaller fragments. Thereby, the molecular compositional parts of the particle contribute to the particle's mass spectrum not only by their own m/z ratio but also by those of their successive sub-fragmentions. For statistical considerations, these fragmentation patterns are constant. Consequently, SPMS data can be described by a latent variable model of independent components that represent fragmentation patterns. It has been shown that, in order to derive the independent components from a non-negative matrix, it is sufficient to find a factorization into a non-negative basis and coefficients for which the coefficients of the basis vectors are uncorrelated [39]. Defining the optimization goal of independence between the basis vector's coefficients, therefore, serves not only the purpose of dimension reduction (basis transformation into independent latent variables) but also leads to basis vectors of distinct molecular composition. Thus, the regularization of independence aims toward an unambiguous and physically interpretable basis.

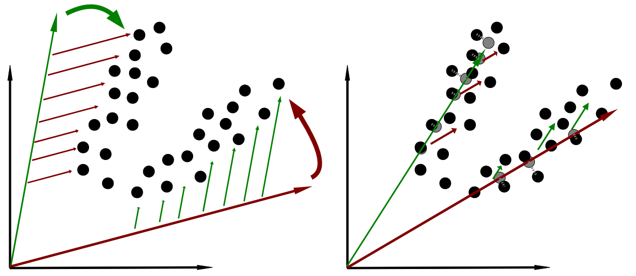

Fig. 2. With regularization, NMF finds a non-negative factorization in coefficients of independent basis vectors (particle components). Thereby, the correctness of the mapping (errors are illustrated in gray) is balanced against the independence and diversity of the basis.

Non-convex optimizations are prone to lead to only locally optimal results. Our experiments of factorizing SPMS data have shown that the objective of independence leads to the fact that outliers are not mapped well with this criterion alone and the optimization of the basis may become "entrapped" by the local solution for independence. Although we present no empirical evidence for this fact, the intuition behind these dilemmas is clear. From an optimization perspective, the gain in correctness by changing the basis toward faces of the bounding box of the data does not outweigh the penalty of correlation induced by this change. Consequently, the optimization terminates in a local optimum. Although independent components of SPMS data have to be *diverse*, early implementations using only the regularization of independence have not produced this result but have ended abruptly in unreliable solutions. However, steering the search for an independent basis by including a slightly weighted regularization of diversity has produced far more reliable results.

To summarize these considerations, the non-negative matrix factorization of SPMS data has the following objectives.

- NMF: C and B define a factorization of X for which $||X - CB||$ is minimized → numerically correct
- Regularization w.r.t. independence: basis coefficients C are decorrelated → unambiguous basis
- Regularization w.r.t. diversity: basis vectors B are mutually different → reliable solution

We use a combination of [39] and [20] that involves a gradient-based two-block optimization scheme with multiplicative update rules according to [23]. The computations can be summarized as follows, where $||.||_F^2$ denotes the squared Frobenius norm.

1. **Numerical correctness** is enforced by multiplicative update rules in two successive blocks:

$$min_{C\geq 0} \qquad ||X - CB||_F^2 \text{ , by} \tag{3}$$

$$C_{a,b} \leftarrow C_{a,b} \frac{\left([XB^T - \alpha_C R_C]_{\geq\varepsilon}\right)_{a,b}}{(CBB^T)_{a,b} + \varepsilon} \text{ ,}$$

where B is fixed and

$$min_{B\geq 0} \qquad ||X - CB||_F^2 \text{, by} \tag{4}$$

$$B_{a,b} \leftarrow B_{a,b} \frac{\left([C^T X - \alpha_B R_B]_{\geq\varepsilon}\right)_{a,b}}{(C^T CB)_{a,b} + \varepsilon} \text{ ,}$$

where C is fixed. Here, $[.]_{\geq\varepsilon}$ denotes that values are truncated to be greater or equal to a small positive real value. The update rules employed here are inherently "normal" additive gradient updates with a relative step size. However, this multiplicative formulation yields faster computations and is currently the dominant approach to NMF [23]. The regularization terms R_C and R_B are weighted by α_C and α_B, respectively, producing an independent and diverse basis.[1] R_C and R_B are the partial derivatives of the cost functions $J_C(C)$ and $J_B(B)$, i.e.,

$$R_C = \frac{\partial J_C(C)}{\partial C} \text{ and} \tag{5}$$

$$R_B = \frac{\partial J_B(B)}{\partial B} \text{ .} \tag{6}$$

[1] Note that α_B should be distinctly lower weighted than α_C.

Detailed notations of R_C can be found, for example, in [39].

2. **Independence** is enforced by updating C using the partial derivative of the cost function $J_C(C)$ that defines the discrepancy between the *uncentered* correlation (normed uncentered covariance) matrix of C, $nCorr(C)$, to the $k \times k$ identity matrix, I_k.

$$\begin{aligned} J_C(C) &= ||\, nCorr(C) \; - \; I_k||_F^2 \text{ , with} \\ nCorr(C) &= N_C C^T C N_C \text{ ,} \\ N_C &= diag(1/||C_{\bullet,1}||_F, ..., 1/||C_{\bullet,k}||_F) \text{ , and} \\ ||C_{\bullet,i}||_F &= \sqrt{\sum_{1 \leq l \leq n} C_{l,i}^2} \text{ .} \end{aligned} \tag{7}$$

3. **Diversity** is enforced by updating B using the partial derivative of the cost function $J_B(B)$ that defines the discrepancy between $cos(B)$, the $k \times k$ matrix of the cosines of the angles between all basis vectors in B, $cos(B)$, to the $k \times k$ identity matrix, I_k.

$$\begin{aligned} J_B(B) &= ||\, cos(B) \; - \; I_k||_F^2 \text{ , with} \\ cos(B) &= N_B B B^T N_B \text{ ,} \\ N_B &= diag(1/||B_{1,\bullet}||_F, ..., 1/||B_{k,\bullet}||_F) \text{ , and} \\ ||B_{i,\bullet}||_F &= \sqrt{\sum_{1 \leq l \leq m} B_{i,l}^2} \text{ .} \end{aligned} \tag{8}$$

Note that computation of the partial derivative of J_B is algorithmically equivalent to that of J_C. Instead of minimizing correlation between columns in C ($C_{\bullet,i}$), R_B maximizes the cosine of the angle between rows (basis vectors) of B ($B_{i,\bullet}$). Figure 3 provides an example of how this NMF implementation behaves in two dimensions.

4.2 Visual encodings

Although the numerical optimization, as presented in the previous section, has desirable mathematical qualities with respect to factorizing mass spectra, results are not guarantied to be physically meaningful to domain experts. Due to noise, outliers, or local optimality, the verification of the solution is required by scientists. In contrast to previous approaches using numerical error metrics, we provide domain and problem-specific visual representations of the matrix factorization process. This enables scientists to analyze the optimization result in full detail, as well as to assess its quality on an abstract level. In the following, we give a detailed account of these visual encodings and discuss their suitability for air quality research.

The factorization of a data set $X \in R^{(n \times m)}$ introduces additional entities that require visual representation to make the user aware of their properties:

- the basis $B \in R^{(k \times m)}$,
- the coefficients $C \in R^{(n \times k)}$,
- the mapping error $X - CB$, as well as its metric sum $||X - CB||_F^2$,
- the correlation of the basis vectors' coefficients $nCorr(C)$, and
- the cosine of the angle between basis vectors $cos(B)$.

However, not all entities are of equal interest (or importance) in the validation of SPMS matrix factorization. Most importantly, B must be visualized, as the physical correctness of the basis defines the value of the factorization. If the basis vectors cannot be interpreted as a meaningful physical entity in the application, the factorization will hold no physical meaning to scientists. Enabling the analysis and validation of B is therefore a key requirement to be met by the visualization. Consequently, the visualization of B must show each basis vector in full detail.

As the basis is to be evaluated in relation to the data, the visualization must also involve the depiction of X in equal detail. This leads to the conclusion that X and B have to be depicted in the same visual space and form, such that the user can visually reference both entities in relation to each other more easily. Both B and X are histogram data for which several alternative visual representations exist. However, mass spectrometry data has already an established visual representation in the engineering community which novel visualizations have to conform to. As Figure 1 shows, mass spectrograms are visually represented as piecewise linear functions over mass. The effectiveness of our method depends on the scientist's ability to investigate patterns in these spectra, as well as on the experience in identifying diverse sources contributing to SPMS data. Therefore, for the histogram data visualization of X and B, the basic geometric form must relate to the existing representation. If X and B are seen as multi-dimensional, then piecewise linear functions equal the representation by parallel coordinates [16] with the following exceptions:

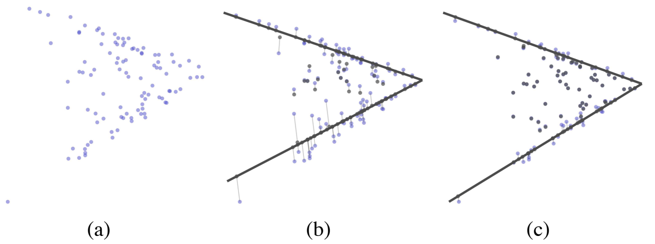

Fig. 3. An example of our NMF implementation. (a) shows a 2D point arrangement that reflects the geometric properties of the SPMS data model as described in Section 3.1, has two fairly independent sources, and noise added to it. Basis vectors and coefficients are computed by steepest descent (b), after which the unknown non-negative mixture of the two contributing sources is found almost exactly (c).

- No vertical lines are drawn for the dimensions.
- Every dimension has equal scale.
- The order of dimensions is not arbitrary but given by mass ratios.

These exceptions render the application of state-of-the-art visualization techniques for parallel coordinates (see Section 2) very limited. For visualizing SPMS data, dimension reordering, edge bundling, or other forms of abstraction are generally less accepted by air quality researchers. However, transfer functions, alpha blending, as well as coloring schemes, that slightly affect the visual appearance but not the spatial presence of data features, are accepted degrees of freedom for the visual representation.

Next to the basis, the mapping error $X - CB$ is of equal importance in the verification of SPMS factorization. Even if the basis is verified to be physically correct, if the factorization does not hold for the data, it will be of little worth to successive analysis. While the overall quality of the factorization can be assessed by the norm $||X - CB||_F^2$, a detailed visualization of $X - CB$ gives clues as to how the basis can possibly be adjusted to achieve a better factorization result. NMF optimization is prone to local minima and the computed basis is most likely not optimal. However, adjustment of B with respect to the mapping error can improve results and lead to finding a global optimum. Consequently, a detailed visualization of the mapping error is equally crucial to the effectiveness of our method.

There are two possibilities for defining the mapping error: absolute ($X - CB$) and relative (CB/X), both are interesting for atmospheric scientists; even small measurements in specific m/z dimensions can be important. As numerical optimization based on least sum of squares tends to neglect small values, the visualization of the relative error, showing the factorized data in relation to the original data, is required for verification. On the other hand, the absolute error is of equal importance as it allows for the assessment of information loss and shows the patterns of features that are not captured in the factorization. Consequently, our visualization has to be able to depict both absolute and relative factorization error in detail, while the absolute error should be displayed in relation to the basis in order to give insights to its adjustment. Therefore, we depict $X - CB$ as separate polylines in the same axis as X and B. To reduce confusion between these three visual entities, we use distinct colors for each of them. We also allow for the option to hide the absolute error in case it should clutter the view.

The relative error should be visually distinguishable from the absolute error and, therefore, have a distinct representation. In our context, it shows the user how the factorization fits to each coordinate of the data, i.e., whether the mapping accounts for higher or lower values with respect to each value in X. To avoid further cluttering the visualization, we exploit the yet unused degree of freedom of color coding X by this error measure. Although this is intuitive to analysts, the drawback is that color may not be used for other visualization goals, for example, to depict different clusters. Here, we use alpha blending to bring out some of the data's structure. To easily distinguish under- and over-representation, and since relative errors CB/X generally show different distributions in these ranges ($[0,1]$ and $[1,2]$), two distinct colors are required for the definition of the color map. For $CB/X = 1$, the user's attention is not needed as no error is present.

To this point, we have established the necessity of a single high-detail visualization showing B relative to X (colored by the CB/X) and $X - CB$. However, for analysis and verification of the factorization, the user also needs to quickly gain an overview of the general mapping quality. This overview should facilitate an abstract comprehension of the overall fit of the factorization and serve as a platform for interaction and navigation. We provide this overview by a linear projection of X, CB, and B, defined by the two principal components of the covariance matrix of X, as they are the orthogonal axes of maximal variance in X. Further, we connect each point in X with its corresponding mapping in CB. In this projection, the user can quickly identify outliers that are not mapped well by the factorization, as well as its general quality.

A third plot accounts for the visual representation of $nCorr(C)$ and $cos(B)$, as independence and diversity are important aspects of the optimization. However, their depiction is not as essential to data analysis as the projection or graph view. As heat map representations of correlation matrices are well-established in the engineering community and are also space efficient, we join both symmetric matrices by their upper and lower triangular half, respectively, and display their values by color coding in gray scale. Note that the basis has no inherent order and reordering of this matrix is semantically possible, however, we find no need to do so as our NMF implementation generally performs well with respect to independence of the basis.

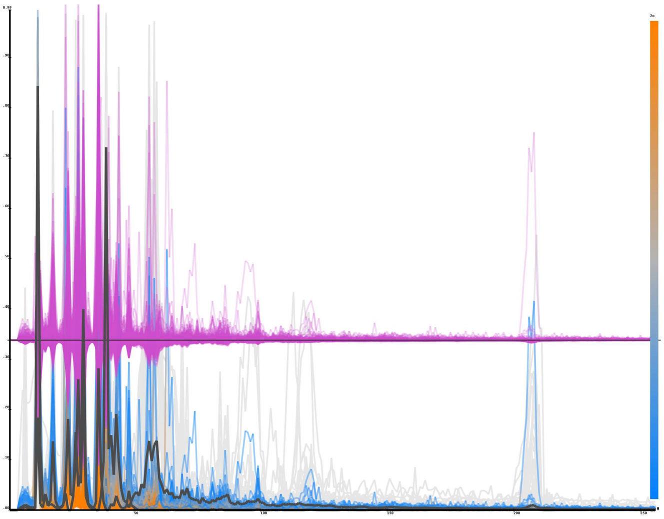

Fig. 4. Overview of particle spectra in 256 dimensions showing absolute (magenta) and relative error (orange-blue) of the computed factorization by two independent basis vectors (dark gray). Spectra are represented as piece-wise linear functions over m/z and exhibit the same patterns as established representations.

4.3 Interaction

Effective visual analysis and manipulation of matrix factorization requires interaction. In this section, we introduce the means for interaction provided by our visual interface and comment on their eligibility for interactive visual verification of SPMS factorization. Figure 5 illustrates these techniques.

As established in previous sections, numerical error measures do not facilitate understanding of complex data features. The visual evaluation of the factorization is a necessary step in air quality research. Thereby, effective interaction techniques are required for atmospheric scientists to analyze the mapping error in each dimension with respect to outliers, noise, and data features. We distinguish between two interaction classes: *analysis* and *manipulation*. For the analysis part, we apply interaction techniques from visual analytics to our visualization of matrix factorization. However, few techniques are available that focus on interfacing dimension reduction methods. Here, we introduce novel and intuitive interaction mechanisms that interface matrix factorization.

The first step in assessing the quality of a factorization result is its visual analysis. Given the level of restrictions for visual encodings, parallel coordinates are inevitably prone to visual clutter with increasing number of data points and dimensions. Therefore, interaction is usually required for effective visual analysis. *Zooming and panning* allows for detailed analysis of specific parts of the factorization. In order to analyze a group of data points or basis vectors, we allow for problem-specific, semantic *selection and filtering* mechanisms. Thereby, the projection plot acts as the selection interface that induces filtering operation in the graph plot. Upon selection, the alpha values of all polylines are adjusted according to their analytic connection to the selection. The selection of points

- hides other points in X,
- hides absolute errors in $X - CB$ not stemming from the selected points, and
- shows basis vectors according to their coefficients in the mapping of the selected points, i.e., $\alpha_b = \sum_{p\ selected} C_{p,b} \ / \ \sum_{p\ selected} ||C_{p,\bullet}||_1$ [2],

while the selection of basis vectors

- hides other basis vectors in B,
- leaves absolute errors in $X - CB$ untouched[3], and
- shows points according to their coefficients in the mapping of the selected basis vectors, i.e., $\alpha_p = C_{p,b} \ / \ ||C_{p,\bullet}||_1$ [4].

After thorough analysis of the factorization and its error, the atmospheric researcher may refine parameters for a successive optimization step, either from an adjusted or random starting point of parameter space. By interacting with the factorization, the analyst can *manipulate the basis* by adjusting coordinates of basis vectors via left-mouse dragging. As basis vectors are normalized by the optimization, the unadjusted coordinates are updated such that the norm holds for the adjustment. By permitting this manipulation, the user can iteratively define new starting points for the gradient-based search of an optimal data basis for factorization.

Non-convex gradient-based optimization can be unpredictable, especially when applied to high-dimensional data. As the definition of

[2] For selected points p, the opacity of each basis b is adjusted to α_b, where $||C_{p,\bullet}||_1$ refers to the sum of the p'th row of the coefficent matrix C.

[3] Note that the mapping errors cannot directly relate to any selection of basis vectors. Highlighting parts of the errors might lead to the semantically incorrect conclusion that the selected basis vectors are accountable for these errors.

[4] For basis vector b selected, the opacitiy of each spectrum (m-D point) p is adjusted to α_p, where $||C_{p,\bullet}||_1$ as above.

different starting points for optimization may not necessarily imply a different result after steepest descent, setting the starting point is not sufficient to assure physical correctness. Additionally, the user can set threshold values for the basis optimization. Relative from the position of each basis coordinate, positive and negative thresholds can be set by right-mouse dragging. These thresholds act as strict boundary limits for the steepest descent which are guaranteed to be met by the optimization. Setting boundary levels for only a few coordinates can change the entire basis in a way that the configuration is optimal regarding to the given restrictions, while the scientist can decide the exact degree of freedom for every part of the basis optimization.

When, after restarting the factorization with a refined basis, the result is still unsatisfactory with respect to the mapping error, the researcher can investigate whether the number of basis vectors is ill-set for the data's factorization by *adding or removing basis vectors*. The complete *randomization* of basis vector coefficients can also provide the necessary means for overcoming local optima. Often, this scenario can be observed while the online visualization is running. Therefore, the optimization can be *stopped or resumed* at any time.

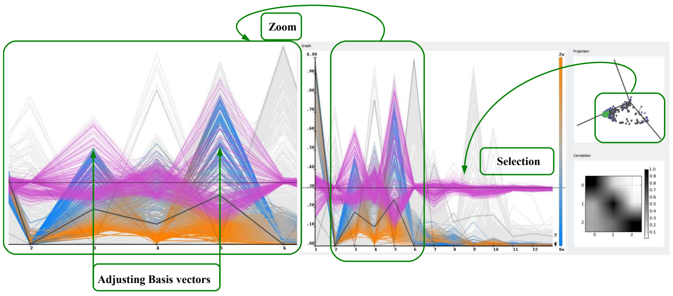

Fig. 5. Using interaction techniques and filtering, atmospheric scientists can analyze the factorization error in specific dimensions, investigate the contribution of basis vectors to the mapping of the spectra. Initial results of the factorization can be adjusted by setting new starting parameters or thresholds for the optimization. Expert verification and steering aims toward physically sensible factorizations and interpretable results.

5 Results

The authors from the domain of air quality research have applied this method to the factorization of biomass combustion sources. Here, we give excerpts of our preliminary study and evaluate the visualization framework[5] with respect to speed, accuracy, and ease of use. We also give a glimpse into future research, however, these findings will be published in a different forum. In summary of what is presented here, we have been able to (i) reproduce established findings in mere a fraction of the time than it was possible before, (ii) process and analyze ten-times more spectra than in previous studies, and (iii) gain surprising insights enabled by the visualization.

5.1 Factorization of biomass combustion sources

Biomass combustion emits copious amounts of gases and particles into the atmosphere and plays a key role in almost all present day environmental concerns including the health effects of air pollution, acid rain, visibility reduction, and stratospheric ozone depletion. Among the largest inadequacies in quantifying emission factors of biomass combustion is the general paucity of methods identifying and quantifying particle classes in ambient measurements for a wide range of ecosystems and combustion conditions, including anything from naturally occurring, large wildfires to woodstoves and fireplaces used for residential heating [31]. This is largely a result of the physical and chemical complexity of particulate matter (PM). PM is the least understood factor in almost all issues ranging from human health to global climate change and provides the impetus for studying ambient particles in increasing detail to close these knowledge gaps.

In prior work, we have used high-resolution clustering algorithms to characterize particle classes of biomass combustion emission factors [3]. The particle class depictions in Figure 6 are the averages of all single particle mass spectra within their class and the listed percentages represent the fraction of the total detected particles belonging to that class. The relevant carbon cluster ions C_x^+ (typically attributed to elemental carbon (EC)), hydrocarbon fragment ions $C_xH_y^+$ (organic carbon (OC)) and isotopic K^+ ions are labeled. However, the apparent distinction between these particle classes, or compositional discretization, is somewhat arbitrary and largely a result of the parameters chosen to control the data clustering algorithm. In reality, there is a "continuous distribution" of particle compositions ranging from those with mass spectra dominated by K (Figure 6.a) to purely carbonaceous aerosol with mass spectra dominated by EC and OC ions (Figure 6.d). The fundamental issue in correctly distinguishing these particle types is isobaric interference at m/z 39. Values in m/z 39 can represent K^+ ions, $C_3H_3^+$ ions or some mixture of the two. $C_3H_3^+$ ions are fairly ubiquitous in the mass spectra of carbonaceous aerosol, and thus our ability to accurately separate biomass combustion from other sources of EC/OC particles in ambient mixtures resides almost exclusively in our ability to accurately determine the relative contribution of these two ions to the signal intensity observed at m/z 39. The progression of particle compositions shown in Figure 6 was designed in attempts to capture this issue. The presence of K is unambiguous in Figure 6.a, and even Figure 6.b. As the ion signal at m/z 39 decreases and becomes comparable to, and eventually less than, the C_3^+ signal (in Figures 6.c and 6.d), it is increasingly difficult to unambiguously specify the presence of K in the particle. This is a predominant issue as Figure 6.d is characteristic of what is generally observed for carbonaceous particles from a variety of sources. As a result, a significant amount of manual effort is expended making the appropriate peak assignments necessary to distinguish between purely EC/OC and biomass combustion particles. Unfortunately, this cannot be done using the cluster-average mass spectra, such as those shown in Figure 6, since the relevant clusters commonly contain a mixture of both particle types and clustering obscures important details in the spectra that assist the analyst when making the assignments. Instead, the individual mass spectra comprising the clusters must be inspected, interpreted and classified manually to obtain more accurate source separations. This is extremely time-consuming and ultimately based on the subjective interpretation of an expert analyst without a fundamental mathematical underpinning.

In the following, we discuss the application of our method to (1) the woodstove source sampling data from Pittsburgh, Pennsylvania, discussed above and (2) ambient Rapid Single-ultrafine-particle Mass Spectrometer (RSMS) data collected during a sampling campaign in Fresno, California [2].

We evaluate this method based upon (i) accuracy in resolving K-containing particles from EC/OC particles, (ii) efficiency in reducing overall analysis time and (iii) the contribution of the visualization and interactive elements in conducting, interpreting, and gleaning new insight from the overall process. The woodstove source sampling experiments offer a good basis to evaluate performance since the data are already well-characterized while the Fresno campaign provides a relatively unexplored and complex ambient data set to test the visualization and interactive elements. The Fresno data is particularly well-suited for this study since the two largest sources of particle pollution in the area are vehicular emissions from local traffic and biomass combustion emissions from residential heating and agricultural burning. As a result, the composition of the air shed is a large external mixture of internally mixed EC-, OC- and K-containing particles, as well as components formed in the atmosphere by gas-phase photochemistry and condensed on the particles, providing a very challenging environ-

[5] The computation speed of NMF strongly varies, depending on the number of data points (n), dimensions (m), and basis vectors (k), as well as the desired quality of the factorization. A factorization of average quality by a prototypical implementation, written in Python using NumPy, runs several minutes for n=1.000, m=256, and k=3, on standard off-the-shelf hardware. For larger data sets, with n >200.000, the NMF routine can take several hours to converge.

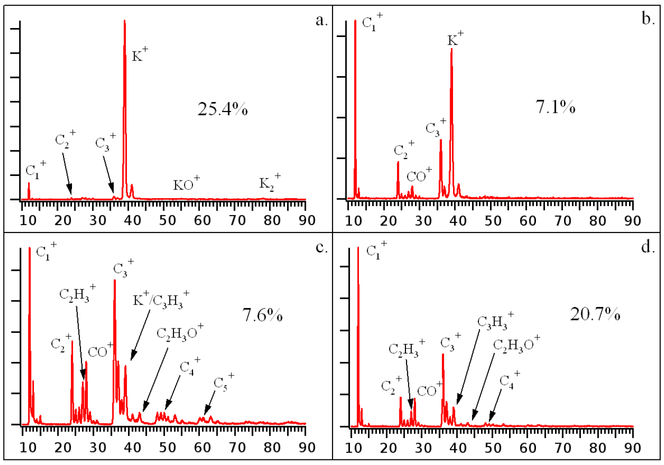

Fig. 6. Single particle mass spectral representations of particle types observed from biomass combustion in a wood stove during source sampling experiments. Due to isobaric interference at m/z 39 (K+ / C3H3+), a clear separation could not be achieved.

ment for resolving sources.

Results from the interactive analysis of the RSMS data collected during the wood stove source sampling experiments are shown in Figure 7. Only those dimensions, or m/z values, relevant to resolving the ambiguity in the presence of particulate K were included in the analysis: m/z 24 (C_2^+), 27 ($C_2H_3^+$), 36 (C_3^+), 39 ($^{39}K^+$ / $C_3H_3^+$), and 41 ($^{41}K^+$ / $C_3H_5^+$). This ability to select dimensions of interest, rather than analyzing the full m/z range of the data, is a strong feature of the visualization interface and reduces both computational burden and analysis time tremendously. The top panel of Figure 7 is a screen shot of the visualization interface showing all of the data (fine lines) and basis vectors (bold lines) identified during the factorization. It is immediately clear in this figure that the algorithm does an excellent job factoring out the K^+ and $C_3^+/C_3H_3^+$ signals and that these two basis vectors model the data well. Also apparent is the fact that the solution is slightly over-determined and the elements of the C_2^+ basis vector could have been incorporated into the C_3^+ basis vector without any loss of information. The major advantage of the visualization interface is that this can be done interactively by removing the C_2^+ basis, adjusting the C_3^+ basis vector lines to match the C_2^+ and $C_2H_3^+$ signal evident in the data and then performing the optimization again. This is a very useful and efficient way of analyzing these data. An unexpected and highly interesting result is the apparent irrelevance of the m/z 41 dimension. Previous efforts have been made to separate biomass combustion and EC/OC particles based on the ratio of integrated ion signal at m/z 39 to 41, and thus its inclusion in the analysis, but with limited success. The underlying assumption is that $C_3^+/C_3H_3^+$ ratios are small compared to the larger values associated with the natural isotopic abundances of ^{39}K and ^{41}K. Further research shall be conducted in this regard.

The bottom left panel of Figure 7 shows the projection of all data points onto the K^+ and C_3^+ basis vector space. Again, the separation in the data is strikingly clear with K dominant particles clustered in the upper left-hand corner and EC/OC dominant particles in the bottom right; these areas are circled in the figure. This interactive visualization framework is very resourceful and the ability to interact with the projection by highlighting individual data points, or clusters of points, and inspecting the relative basis contributions to the selected points, is invaluable to interpreting the results and understanding the structure of the data.

To quantitatively separate K-containing particles from purely EC/OC particles, the C_3^+ basis vector was factored out of all data points and the resulting data re-projected onto the K^+ and C_3^+ basis vector space, as shown in the bottom right panel of Figure 7. The idea is that those data points showing near-zero contribution from the K^+ basis vector are purely EC/OC particles while the above-zero points are K-containing particles. Clearly, there is error in the overall fit of the basis vectors to the data, as evidenced by points with negative K^+ contribution, and this must be incorporated into the analysis. Using the average residuals between the data points and basis vectors at m/z 39 shown as bright pink lines in the top panel of Figure 7 as an error estimate for the factorization, a window centered about zero has been drawn in the projection designed to separate purely EC/OC particles from K-containing particles; note that the window fully encompasses those points with negative contributions. Summing all points within the window yields a value of 0.29 for the fraction of the total number of particles sampled that are purely EC/OC particles. This result is in nearly perfect agreement with the value of 0.32 obtained during the manual analysis of all 7000 mass spectra [3].

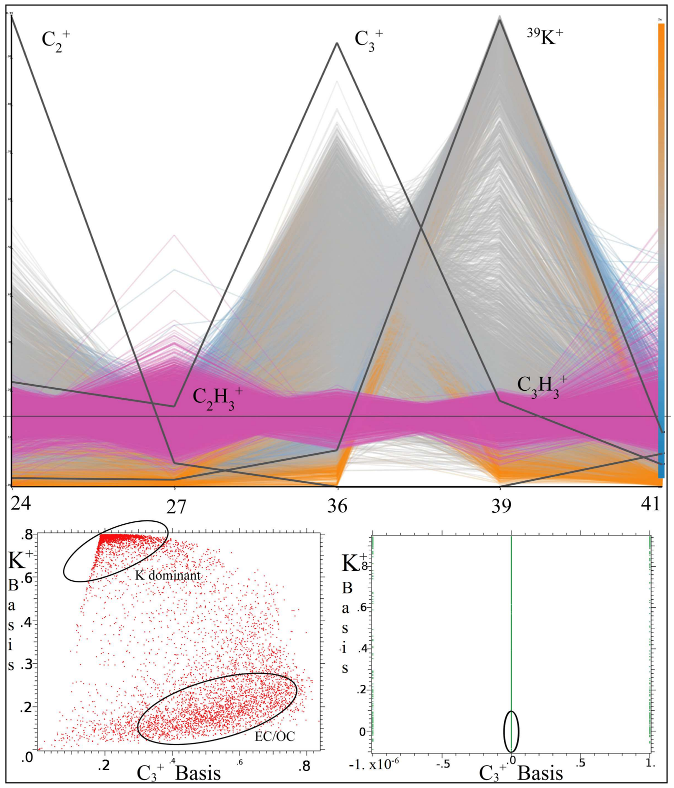

Fig. 7. Top: Screen shot of the visualization interface showing results from factorization of the RSMS data collected during the PAQS woodstove source sampling experiments. Bottom left: Projection of the RSMS data onto the K^+ and C_3^+ basis space identified during factor analysis. Bottom right: Projection of the RSMS data onto the K^+ and C_3^+ basis space after factoring out the C_3^+ basis vector from the data; As can be seen in the top figure, the focus for dimension reduction lies with information in C_2^+, C_3^+, and K^+, and not in $C_2H_3^+$ or $C_3H_5^+$. The factorization minimizes the error in mapping K^+ and C_3^+ signals which corresponds nicely to existing research stating that K^+ is the major classifier in these dimensions. Further, the spectra of K^+ and $C_3^+/C_3H_3^+$ particle classes are separated automatically and accurate. In prior work, atmospheric scientists have required months to obtain this result.

Figure 8 shows results from the interactive analysis of the RSMS data collected during the field campaign in Fresno. The same five dimensions were used and over 70,000 single particle mass spectra were analyzed. Quite notably, this would not have been possible by manual analysis. The optimization was initiated with three basis vectors but interactively reconfigured to only two during analysis. The projection of all data onto the K^+ and C_3^+ basis vector space is shown in the lower left panel and the clustering of the data points is very distinct and clear. A majority of the data appears to fall roughly along a positively sloping line (circled in the figure), where the upper cluster represents K-dominant particles and the K^+ contribution decreases down the line toward the lower cluster. A snap shot of the visualization interface when the lower cluster of data points is highlighted is shown in the top panel. The apparent weak contribution of both the K^+ and C_3^+ basis vectors to this cluster is actually due to the prevalence of NO^+ ions (m/z 30) in these mass spectra. The factorization and visualization features do an excellent job resolving this particle class, especially since m/z 30 was not included in the analysis. Purely EC/OC particles cluster in the lower right-hand corner of the projection but are very sparse relative to the other particle types, and even the results of the woodstove source sampling. This is also a highly interesting and informative result but will not be addressed any further here. Similar to the analysis above, the C_3^+ basis vector was factored out of all data points and the resulting data re-projected onto the K^+ and C_3^+ basis vector space (lower right panel). Again, an error estimate based on residuals was used to create a threshold range about zero to calculate a value of 0.06 for the fraction of the total number of particles sampled that are purely EC/OC particles. A notable strength of this particular interactive exercise was the ability to robustly differentiate the presence of K^+ versus $C_3H_3^+$ even in spectra where NO^+ dominates the ion signal.

5.2 Expert evaluation

When processing high-dimensional data, paying attention to all the dimensions is challenging, so it is crucial to provide the user with a mechanism for appropriately reducing the dimension of the problem at a minimum loss of information, as well as showing the user both which dimensions are important and where information is lost. In the example here, subtle differences between mass spectra can provide crucial guidance for assessing and quantifying the source contributions to a given air shed. Clustering algorithms may obscure these subtleties reducing their usefulness. At some level, the involvement of an expert analyst is unavoidable but the burden of manually inspecting the several hundred thousand mass spectra acquired during typical field campaigns, or even a subset thereof, is unreasonable, time-consuming, and largely subjective. A highly visual and interactive computational platform for analyzing, characterizing and manipulating these data is essential to these efforts. The emerging data visualization community requires interdisciplinary collaborations to develop effective platforms – the impetus for the current work.

The interactive visual framework developed here provides educated, mathematically rigorous suggestions, while leaving full control and physical verification to the analyst. In this regard, both visualization and interaction build trust in both the factorization method and its implementation. These computations may identify large errors, which must be conveyed to the analyst. While information loss may be unavoidable at some level, it is crucial to visualize exactly where this loss occurs. Both the color coding and the error lines are helpful in qualitatively interpreting the basis transformations related to potential errors. Filtering spectra and basis vectors is intuitive and by merely performing a few interactions, the dominant clusters of spectra are easily highlighted. This astonishing level of visual interaction with mass spectrometry data has not been possible before, thereby introducing both new and exciting research potentials.

While previous methods did not recognize important features in the spectra, the factorization is much more facile at identifying important commonalities and subtleties in the spectra. We were able to reproduce established findings from the woodstove biomass combustion measurements in a matter of hours, where prior work took months. Also we were able to gain new insights from data collected in Fresno that will be the focus of future investigations. The factorized data is lower-dimensional and resolves isobaric interferences verifiably so gives a considerably better starting point for subsequent data processing, analysis, and visualization.

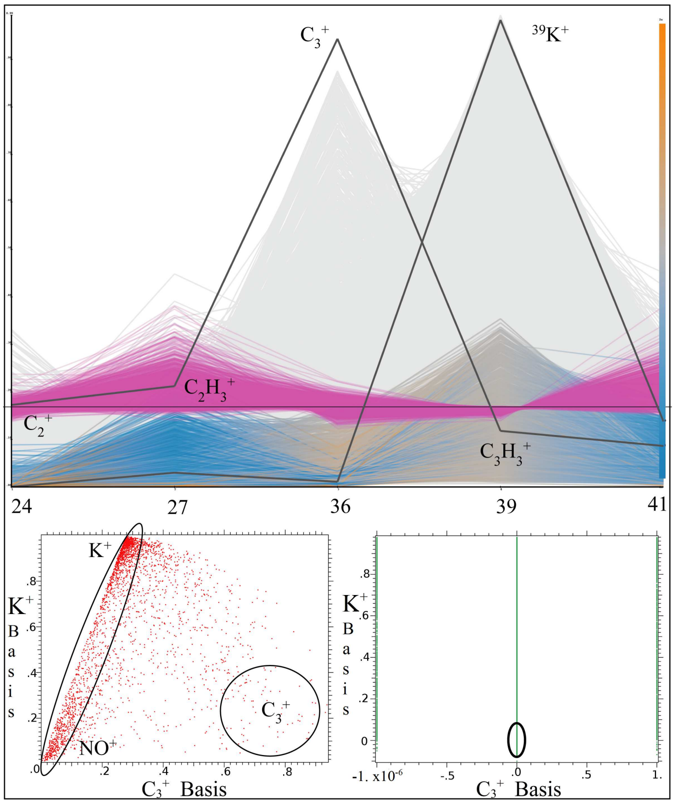

Fig. 8. Top: Screen shot of the visualization interface showing results from factorization of the RSMS data collected during the field campaign in Fresno, CA. Bottom left: Projection of the RSMS data onto the K^+ and C_3^+ basis space identified during factor analysis. Bottom right: Projection of the RSMS data onto the K^+ and C_3^+ basis space after factoring out the C_3^+ basis vector from the data; While K^+ and C_3^+ are captured and separated well in the factorization, the projection by the basis suggests a dominant cluster in the data that is not part of these two particle classes. Further analysis shows that this is due to the prevalence of NO^+ ions (m/z 30) in these mass spectra.

6 Conclusion and Future Work

We have presented a framework for dimension reduction of single particle mass spectrometry data that entails the use of visualization and interaction in order to steer computations. Our work contributes to the community of air quality research by providing novel means to reduce data dimensionality by physically correct and unambiguous basis transformation. Thereby, we overcome limitations of previous methods that use dimension reduction as a black-box scheme and move toward more physically sensible computing. By the utilization of specifically tailored regularization terms, the presented non-negative matrix factorization is capable of resolving ambiguity of the mass spectrum, thereby, laying the groundwork for more reliable data clustering in air quality research. As the expert evaluation shows, analysts are able to reproduce known research results with great ease and speed by using our method. To the visualization community, we contribute well-justified visual encodings, as well as novel interaction techniques that aid in the visual analysis and verification of matrix factorization. Future work work will be directed at improving the visual representation of SPMS data by applying clutter reduction techniques from the field of information visualization, as well as improving the runtime of matrix factorization by large-scale parallel tensor multiplication.

Acknowledgments

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), as a specific project of the International Research Training Group (IRTG) 1131.

References

[1] A. O. Artero and M. C. F. de Oliveira. Levkowitz h.: Uncovering clusters in crowded parallel coordinates visualizations. *IEEE Symp. on Information Visualization*, 2004.

[2] K. J. Bein, Y. Zhao, and A. S. Wexler. Conditional sampling for source-oriented toxicological studies using a single particle mass spectrometer. *Environmental science technology*, 43(24):9445–9452, 2009.

[3] K. J. Bein, Y. J. Zhao, A. S. Wexler, and M. V. Johnston. Speciation of size-resolved individual ultrafine particles in pittsburgh, pennsylvania. *Journal of Geophysical Research*, 110(D7):1–22, 2005.

[4] D. Carlsen, M. Malone, J. Kollat, and T. W. Simpson. Evaluating the performance of visual steering commands for user-guided pareto frontier sampling during trade space exploration. *ASME Conference Proceedings*, 2008(43253):499–509, 2008.

[5] D. Chen and R. J. Plemmons. Nonnegativity constraints in numerical analysis historical comments on enforcing nonnegativity. *Office*, pages 1–32, 2007.

[6] P. Comon. Independent component analysis, a new concept? *Signal Process.*, 36(3):287–314, Apr. 1994.

[7] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 1st edition, 2010.

[8] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14:1141–1148, 2008.

[9] E. Esser, M. Möller, S. Osher, G. Sapiro, and J. Xin. A convex model for non-negative matrix factorization and dimensionality reduction on physical space. *Arxiv preprint arXiv11020844*, stat.ML:14, 2011.

[10] B. J. Ferdosi and J. B. T. M. Roerdink. Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. *Comput. Graph. Forum*, 30(3):1121–1130, 2011.

[11] G. Grinstein, M. Trutschl, and U. Cvek. High-dimensional visualizations. In *Proceedings of Visual Data Mining workshop, KDD'2001*, 2001.

[12] D. S. Gross, R. Atlas, J. Rzeszotarski, E. Turetsky, J. Christensen, S. Benzaid, J. Olson, T. Smith, L. Steinberg, J. Sulman, A. Ritz, B. Anderson, C. Nelson, D. R. Musicant, L. Chen, D. C. Snyder, and J. J. Schauer. Environmental chemistry through intelligent atmospheric data analysis. *Environmental Modelling & Software*, 25(6):760 – 769, 2010.

[13] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, page 127, Washington, DC, USA, 2002. IEEE Computer Society.

[14] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[15] S. Ingram, T. Munzner, and M. Olano. Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics*, 15:249–261, 2009.

[16] A. Inselberg. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2009.

[17] H. Jänicke, M. Böttinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 14:1459–1466, 2008.

[18] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure within clustered parallel coordinates displays. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, pages 17–, Washington, DC, USA, 2005. IEEE Computer Society.

[19] E. Kim, S. G. Brown, H. R. Hafner, and P. K. Hopke. Characterization of non-methane volatile organic compounds sources in houston during 2001 using positive matrix factorization. *Atmospheric Environment*, 39(32):5934–5946, 2005.

[20] J. Kim and H. Park. Fast nonnegative matrix factorization: an active-set-like method and comparisons. *Science*, 2008.

[21] O. Kreylos, A. M. Tesdall, B. Hamann, J. K. Hunter, and K. I. Joy. *Interactive Visualization and Steering of CFD Simulations*, pages 25–34. Eurographics Association, 2002.

[22] R. S. Laramee, C. Garth, H. Doleisch, J. Schneider, H. Hauser, and H. Hagen. Visual analysis and exploration of fluid flow in a cooling jacket. In *In Proceedings IEEE Visualization 2005*, pages 623–630, 2005.

[23] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

[24] K. T. McDonnell and K. Mueller. Illustrative parallel coordinates. *IEEE-VGTC Symposium on Visualization 2008*, 2008.

[25] A. Messac and X. Chen. Visualizing the optimization process in real-time using physical programming. *ENGINEERING OPTIMIZATION*, 32(6):721–747, 2000.

[26] D. M. Murphy, A. M. Middlebrook, and M. Warshawsky. Cluster analysis of data from the particle analysis by laser mass spectrometry (palms) instrument. *Aerosol Science and Technology*, 37(4):382–391, 2003.

[27] P. Oesterling, C. Heine, H. Jänicke, and G. Scheuermann. Visual analysis of high dimensional point clouds using topological landscapes. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE*, pages 113 –120, Mar. 2010.

[28] F. Paulovich, D. Eler, J. Poco, C. Botha, R. Minghim, and L. Nonato. Piece wise laplacian-based projection for interactive data exploration and organization. *Computer Graphics Forum*, 30(3):1091–1100, 2011.

[29] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[30] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *In INFOVIS 04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS04*, pages 89–96. IEEE Computer Society, 2004.

[31] J. S. Reid, R. Koppmann, T. F. Eck, and D. P. Eleuterio. A review of biomass burning emissions part ii: intensive physical properties of biomass burning particles. *Atmospheric Chemistry and Physics*, 5(3):799–825, 2005.

[32] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[33] X.-H. Song, P. K. Hopke, D. P. Fergenson, and K. A. Prather. Classification of single particles analyzed by atofms using an artificial neural network, art-2a. *Analytical Chemistry*, 71(4):860–865, 1999.

[34] G. Stump, S. Lego, M. Yukish, T. W. Simpson, and J. A. Donndelinger. Visual steering commands for trade space exploration: User-guided sampling with example. *Journal of Computing and Information Science in Engineering*, 9(4):044501, 2009.

[35] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[36] W. Torgerson. *Theory and methods of scaling*. Wiley, 1958.

[37] M. O. Ward, G. Grinstein, and D. A. Keim. *Interactive Data Visualization: Foundations, Techniques, and Application*. A. K. Peters, Ltd, 2010.

[38] J. Waser, R. Fuchs, H. Ribicic, B. Schindler, G. Blöschl, and E. Gröller. World lines. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1458 –1467, nov.-dec. 2010.

[39] K. W. Wilson and B. Raj. Spectrogram dimensionality reduction with independence constraints. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1938 –1941, march 2010.

[40] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symposium on Information Visualization*, 2003.

[41] A. Zelenyuk, D. Imre, Y. Cai, K. Mueller, Y. Han, and P. Imrich. Spectraminer, an interactive data mining and visualization software for single particle mass spectroscopy: A laboratory test case. *International Journal of Mass Spectrometry*, 258(13):58 – 73, 2006.

[42] A. Zelenyuk, D. Imre, E. J. Nam, Y. Han, and K. Mueller. Clustersculptor: Software for expert-steered classification of single particle mass spectra. *International Journal of Mass Spectrometry*, 275(13):1 – 10, 2008.

[43] W. Zhao, P. K. Hopke, and K. A. Prather. Comparison of two cluster analysis methods using single particle mass spectra. *Atmospheric Environment*, 42(5):881–892, 2008.

[44] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual Clustering in Parallel Coordinates. *Computer Graphics Forum*, 27(3):1047–1054, May 2008.