Eurographics Conference on Visualization (EuroVis) 2013
B. Preim, P. Rheingans, and H. Theisel
(Guest Editors)

Volume 32 (2013), Number 3

Towards High-dimensional Data Analysis in Air Quality Research

Submission #: 257, type: application

Abstract

The analysis of aerosol emission sources involves mass spectrometry data factorization, an approximation of high-dimensional data in lower-dimensional space. The optimization problem associated with this analysis is non-convex and cannot be solved optimally with currently known algorithms, resulting in factorizations with crude approximation errors that are non-accessible to scientists. We describe a new methodology for user-guided error-aware data factorization that diminishes this problem. Based on a novel formulation of factorization basis suitability and an effective combination of visualization techniques, we provide means for the visual analysis of factorization quality and local refinement of factorizations with respect to minimizing approximation errors. A case study and domain-expert evaluation by collaborating atmospheric scientists shows that our method communicates errors of numerical optimization effectively and admits the computation of high-quality data factorizations in a simple way.

Categories and Subject Descriptors (according to ACM CCS): I.5.5 [Pattern Recognition]: Design Methodology—Feature evaluation and selection

1. Introduction

Atmospheric particles have been shown to increase morbidity and mortality in urban areas and to alter the Earth's radiative energy balance. A key step in delineating this problem is identifying the emission sources of ambient airborne particles. Using innovative instruments, atmospheric scientists are now able to chemically analyze aerosols in real time, providing unprecedentedly rich data sets for air quality research. These single particle mass spectrometers (SPMS) measure the mass spectrum of aerosols, thereby, fundamentally characterizing particles in high-dimensional space. An exemplary mass spectrum is shown by Figure 1. In order to factor out emission sources from these measurements, analysis requires non-negative matrix factorization (NMF). The optimization problem can be defined as follows: given data that is derived from a combination of unknown sources in unknown abundance and combination, the goal is to factor out both unknowns, provided only with an estimate of the number of sources and an assumption of their mixing model. In air quality research, sources represent a non-negative (and non-orthogonal) basis in high-dimensional space, by which SPMS samples are approximated linearly as coefficients to the basis. However, computing suitable basis vectors and coefficients proves difficult in practice, as the optimization problem is ill-posed and non-convex. Currently known algorithms produce sub-optimal factorization results. The approximation error can be defined as the discrepancy between data and its lower-dimensional approximation. While such errors are, in general, unavoidable in dimension reduction, they can be increasingly large for sub-optimal factorizations and hard to assess by atmospheric scientists without the proper visual analytical tools. However, the visual communication of errors in non-negative matrix factorization has not been studied in visualization research and common visualization tools are not applicable to this problem.

We discuss our new approaches for the visual analysis of approximation errors in non-negative matrix factorization, by describing (i) a methodology to assessing the quality of a factorization basis based on the amount of information introduced by each basis vector, (ii) a visualization of factorization errors designed to depict the major features that are in the data but not included in its factorization, and (iii) means to interactively minimize specific errors. During analysis, the scientist can compare the numerical benefit in introducing a basis vector that minimizes error features selected in the visualization against the benefit of each vector currently in the basis. Following this methodology, the scientist can discover and overcome "being stuck" in local optima of non-convex factorization interactively, consequently improving the factorization quality. Due to the high degree of interac-

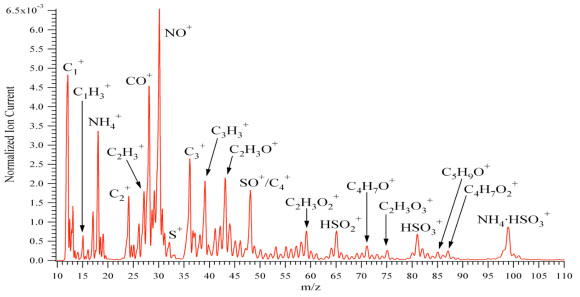


Figure 1: *The mass spectrum of an aerosol represents a pattern (coordinates) that quantifies the abundance of inherent fragment ions (peak labels) per mass (dimensions). Data factorization provides lower-dimensional representations of aerosols in terms of latent components of these patterns.*

tivity in this analysis, our method also provides an awareness about the information loss associated with the dimension reduction process and allows for an educated decision on the degree of freedom needed to approximate high-dimensional SPMS data. Thereby, we contribute both to air quality research by providing novel means that aid in the research of aerosol emission sources, as well as to the science of visual data analysis itself by furthering research on error-aware dimension reduction.

The remainder of the paper is structured as follows. Section 2 discusses related work in data factorization, visualization, and air quality research, while Section 3 provides the necessary background for our effort. Section 4 describes our method, entailing a description of our methodology, the projection of factorization errors, our approach to interactive analysis and refinement of the factorization, as well as implementation remarks. Section 5 demonstrates how this method is effectively applied in the factorization of SPMS data collected during biomass combustion 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

In air quality research, non-negative matrix factorization (NMF) methods are used to classify airborne particle types [KBHH05]. NMF [CJ10] computes a non-negative linear basis transformation that approximates high-dimensional data in lower-dimensional space. As opposed to classical data mining approaches [ZIN*08], NMF is potentially more suitable to support in the interpretation of data from single particle mass spectrometers (SPMS), as it provides non-binary classification of data in terms of non-negative combination of latent physical components. Like many other physical variables and combination, mass is non-negative, rendering non-negativity an integral property for analyzing SPMS data. The NMF method analyzed in this work is based on the original research discussed in [KP08] and [WR10]. 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. Section 4.4 describes a computationally more efficient formulation of the algorithm. A common problem with the approaches mentioned above is that they minimize a non-convex objective function and consequently suffer from the presence of local optima. Other work offers a convex model to NMF but is constrained to a convex combination of data points [EMO*11]. In addition to finding an optimal solution, interpretability is often the greatest problem when working with dimension reduction. Making these approaches more accessible to domain scientists is an ongoing visualization research problem.

In the field of visualization, visual steering of exploration [SLY*09] and simulations [LGD*05, WFR*10] has become a well-established research areas. Enabling user interaction in dimension reduction has demonstrated similar success [PEP*11] and proven that user-guided approaches in data analysis can excel unsupervised methods in terms of quality and interpretability. However, visually interfacing practical engineering optimization has not been a focus of visualization research. Although, visualizing high-dimensional data factorizations can be regarded as part of multivariate data visualization [WGK10, GTC01]. Driven by applications, research focuses on better representation of specific data properties (e.g., scientific point cloud data [OHJS10]), better incorporation of domain-appropriate analysis techniques (like brushing and filtering [JBS08]), or computational speed gains [IMO09]. Other research in this area has focused on enhanced cluster visualization [JLJC05, RZH12], brushing techniques [EDF08, HLD02], abstraction [WBP07], and clutter reduction through dimension ordering [YPWR03, FR11] to enable data comprehension. However, due to the high complexity and dimensionality of SPMS data, as well as the fixed order of dimensions in the mass spectrum, many approaches as, for example, clustering, transfer functions, dimension reordering, or edge-bundling, are not feasible for the visualization of factorization errors.

Recent work [EGG*12] demonstrates that SPMS data analysis can greatly benefit from visualization. Factorization errors were visualized by depicting residuals for every data point in every dimension. An example is given in Figure 2. This visualization is based on parallel coordinates [Ins09] and can lack the capacity for conveying approximation errors effectively. The presentation can become highly dense and cluttered, rendering it unsuitable to analyze factorizations of large data sets. In contrast, the present work focuses on visualizing factorization errors by a projection designed to convey an overview of approximation errors by severeness, type, and abundance. We provide this overview in addition to detailed representations and describe a complete methodology to SPMS factorization analysis.

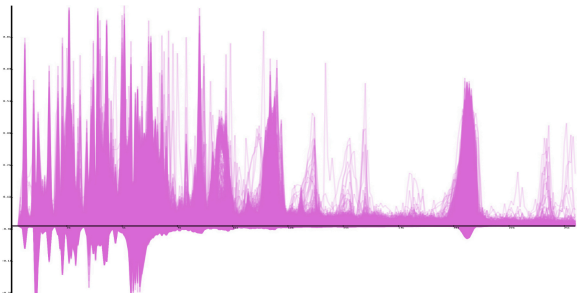


Figure 2: *Previous work visualizes the errors produced by SPMS data factorization in high detail. Due to data complexity and dimensionality, this representation is prone to visual clutter and fails to provide an overview to analysts who are faced with the problem of identifying, classifying, and analyzing error features.*

3. Requirements Analysis

In the following, a brief account of the application background is given. This is followed by a problem definition that involves a description of errors in SPMS factorization and our terminology used in this paper. Finally, we describe the tasks and requirements arising from this problem for the application of air quality research.

3.1. Application background

Single particle mass spectrometry (SPMS) is used in air quality research to categorize, collect, and analyze aerosols at sampling sites of atmospheric interest. Analyzing the particle's composition is an essential step in this research. One way to ascertain a footprint of an aerosol is by single particle mass spectrometry. The mass spectrum of a particle represents a function that maps mass to abundance within the particle. More precisely, it maps the abundance of fragment ions per *mass over elemental charge (m/z)*. Discretized in bins of 1 m/z step size, typical SPMS analyzers capture the first 256 m/z ratios for aerosols. The histogram data is stored as a 256-dimensional positional 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.

Particle composition can be described by the linear combination of latent sub-fragments. Consequently, SPMS data $X \in \mathbb{R}_+{}^{(n\times m)}$, holding n particle spectra discretized in m dimensions, can be described by the m-dimensional mass spectra of fragment ions as a basis B to X, such that

$$X = CB + N \tag{1}$$

Here, B is the matrix storing (row-wise) basis vectors, $B_{j,\bullet} \in \mathbb{R}_+^m$, $1 \le j \le k$, such that X is derived with the coefficient matrix C and the noise N induced by the instrument. Note that all coordinates are non-negative. The problem is ill-posed because C, B, and N, as well as k are unknown, rendering the factorization of SPMS data by an independent basis inherently non-convex. However, the method that is described in the following can cope with these conditions and produce viable solutions to the problem.

Non-negative matrix factorization (NMF) computes a basis $B \in \mathbb{R}_+{}^{(k\times m)}$ and coefficients $C \in \mathbb{R}_+{}^{(n\times k)}$, by minimizing the global mapping error,

$$J = ||X - CB||_F^2 \;\rightarrow\; \min, \tag{2}$$

subject to all values in C and B being non-negative. $||.||_F$ denotes the Frobenius norm. The dominant approach for minimizing J is by updating C and B at each position by its gradient. We apply a gradient-based two-block optimization scheme according to [KP08] and use multiplicative update rules described in [LS00]. We note that minimizing one matrix, while the other is fix, represents a convex optimization problem. We first update C while keeping B fix. If B is initially globally optimal, then updates converge to equally optimal coefficients.

In addition to minimizing the overall mapping error J, applications may impose additional criteria, one of the most common is feature independence. In the context of mass spectrometry, this criterion is understood as the goal of mutually decorrelating the coefficients of basis vectors, which is described by the objective function J_C defined by the squared Frobenius norm of the uncentered correlation matrix:

$$J_C = \sum_{1\le i,j\le k} \left(\frac{(C^T C)_{i,j}}{||C_{\bullet,i}||_F \; ||C_{\bullet,j}||_F} \right)^2 \;\rightarrow\; \min. \tag{3}$$

Thereby, the partial derivative of J_C is evaluated at each position of C for each update. Although this approach to NMF is both flexible and powerful, given the complexity of the problem, drawbacks lie with the slow convergence speed of gradient descent and in its proneness to become "stuck" in local optima. In addition, it requires one to determine an adequate estimate of the number of sources k given as an input. While computational speed can be improved by a GPU implementation, as described in Section 4.4, the latter two problems can most likely not be solved algorithmically. We contribute to solving these problems by describing an error-based methodology to interactive factorization analysis that aids scientists both in uncovering local optima and in making an educated decision concerning the value of k, the number of basis vectors.

3.2. Errors in SPMS data factorization

A multitude of errors are involved in the various stages prior to SPMS data analysis, including but not limited to, data acquisition, sensor measurements, bit noise, integration of the mass spectra, dimension reduction, gradient descent, and visual mapping. While many of these errors are marginal or cannot be determined, the errors introduced by dimension reduction can be both considerably large and determined based

on the original data as ground truth. Given the complexity of high-dimensional SPMS data (that is almost of complete rank), any mapping to lower-dimensional space produces information loss. Although this is inherent to dimension reduction and can be accepted in many domains, it is of particular importance to researchers that rely on non-convex factorization. As opposed to convex problems (for example, singular value decomposition), optimal results of non-convex factorization cannot be achieved by algorithms in realistic time. Moreover, there is no way to ascertain how close a solution is to being optimal.

We focus on visualizing approximation errors introduced by non-convex factorization and further provide means for analyzing the necessity of these errors. Consider a factorization for n data points of dimension m, $X \in \mathbb{R}_+{}^{(n\times m)}$, in coefficients $C \in \mathbb{R}_+{}^{(n\times k)}$ and basis $B \in \mathbb{R}_+{}^{(k\times m)}$ for $k \ll m$. For the purpose of this work, we define *the error of a factorization* as the discrepancy between the original data and its factorization: $X - CB \in \mathbb{R}^{(n\times m)}$. Hence, errors are high-dimensional residuals, given by the misfit for each point in the data. Note that we impose no restrictions on the errors, as they may be both positive or negative and of arbitrary magnitude, as depicted in Figure 2.

In addition to the errors introduced by dimension reduction, a SPMS factorization largely exhibits noise that is assumed to follow a Gaussian distribution (for example, due to gradient descent optimization and sensory noise). For the analysis of a suitable factorization basis, these error contributions are of relatively low interest to analysts, as they are both unavoidable and practically independent of the factorization basis. In contrast, specific error features that are of interest to analysts are those that significantly deviate from a Gaussian distribution. If these specific error features occur in abundance, they strongly indicate that the factorization basis does not allow the depiction of these features in the data. This may be either due to the dimensionality of the basis being set too low, or due to a sub-optimal factorization basis that does not cover significant parts of the data.

In this paper, we make use of terms as significance and optimality. However, it should be noted that optimality, with respect to the analytical purpose of the analyst, can hardly be pre-defined. We resort to this terminology with respect to the quantity of information (variance), as the quality of information cannot be assessed numerically. As such, we define the overall error of a factorization by a norm of its errors ($||X - CB||$) and define a factorization to be optimal that produces a minimal overall error. However, at no point during analysis do we dismiss any solution because of numerical inefficiency. To determine what may serve as adequate to the current purpose of analysis is left to the analyst.

3.3. Requirements and Tasks

Based on our collaboration involving atmospheric and computer scientists, we can conclude that a methodology is needed to (i) assess factorization quality, i.e., the efficiency of a basis in approximating the data, and (ii) assess the errors of a factorization, i.e., the information loss in the approximation. Thereby, analytics to ascertain basis efficiency must be tightly coupled with the visualization of error features (and their significance) to aid the scientist when deciding which errors to admit as a consequence of dimension reduction in order to weight quality against dimensionality of the approximation. As errors may be unnecessarily large for inefficient bases, it is only by the conveyance of both properties (efficiency and error) that scientists can determine the "right" dimensionality for the basis and, consequently, determine an adequate approximation of the data. Finally, this methodology to error-based analysis should include the means to systematically refine factorizations towards minimizing errors. In summary, the key ***tasks and requirements*** for the visual analysis of errors in SPMS data factorizations are:

1. **Analyzing basis efficiency:**
   In assessing the quality of a factorization, it is important to understand where errors originate from, as they may stem from either (i) due to shortcomings of the optimization process (local minima) or (ii) due to a necessity in dimension reduction defined by basis dimensionality. Visualization should help to answer this question and, if possible, uncover inefficiencies of the factorization basis with respect to approximating the data.
2. **Visualizing error features:**
   In dimension reduction, even an optimal factorization basis produces approximation errors. However, the errors that are numerically less important may be more important to the scientist. In order to determine an adequate basis dimensionality and to verify factorization quality, visual assessment of factorization errors is a crucial step. A visualization of factorization errors should convey a classification of errors by importance and type, and serve as a basis to conduct detailed analysis. One major requirement for this is the visual separation of noise from specific error features. Most importantly, the visualization should account for an intuitive assessment of how much of the data is factorized with (less significant) small errors following a normal distribution over all dimensions, as opposed to how much of the data is not well represented, producing errors of (significant) specific features.
3. **Refining factorizations:**
   Once errors are identified during the analysis that are unacceptable, an analytical system should entail the refinement of the factorization towards eliminating these errors. A key necessity of this requirement is interactivity of the data factorization and providing visual feedback concerning the benefit of adjustments.

Our method aims at satisfying these requirements.

4. Method

Since optimization methods cannot guarantee optimal solutions to non-convex problems in practical time, it is hoped that by combining the knowledge and intuition of domain experts with the computing power of machines, admissible solutions can be found. The core of our approach involves therefore interfacing optimization and visualizing factorization errors in a comprehensible manner that allows for analysis and interpretability. Essential to this concept is a highly visual and analytical framework that involves the analyst in several key steps of the factorization. We describe how factorization quality can be analyzed, sub-optimality assessed and the factorization be improved.

4.1. Assessing Optimality

Visualizing optimality of a factorization is a challenging task, as there exists no method that can spot local minima or quantify their (sub-)optimality effectively. However, considering the following concept leads to the conclusion that local minima in non-negative matrix factorization can in fact be revealed with the help of visualization and interaction.

An optimal data basis must consist of basis vectors that are all optimal. Consequently, the exchange of one vector in the basis set must not produce a (numerically) better approximation. Further, for a sub-optimal data basis must hold that there are better basis vectors. These would contribute less to the overall error, with respect to lowering error magnitudes in their abundance (accumulation) (see ()). Moreover, the more errors are similar, with respect to collinearity, the better they can be approximated by a basis vector. Thus, the numerical benefit of a vector to be included in the basis, is directly reflected by and can be identified based on similar errors of high magnitude and abundance. This leads one to conclude that optimality can be assessed by comparing the amount of information that can be conveyed by a basis vector candidate hidden in the error against that of each vector currently in the basis. Consequently, visually highlighting error magnitudes, similarity, and abundance, as well as introducing a measure of information content per basis vector and accounting for visual comparison between each basis vector's numerical benefit, local minima in the factorization can be revealed in analysis.

Based on these considerations, we compute and visualize a "gain" measure for each basis vector that quantifies how each individual basis vector (and its coefficients) contributes to the reconstruction of the data. By visualizing the gain of each basis vector, both in relation to each other, as well as in relation to potential basis candidates (selected by the user based on errors to eliminate), we can visually expose sub-optimal solutions. Since sub-optimal solutions for $b \in \mathbb{R}^m$ and $C_b \in \mathbb{R}^n$ contribute little to large parts of X, they produce (a) a small gain and (b) large errors. Thereby, spotting a local optimum reduces to identifying the basis vectors of small gain and comparing them to the gain of the basis vectors that eliminate large errors. This requires two concepts: (1) visualizing the gain of each basis vector and (2) visualizing the benefit after adjustment of the basis (as direct visual feedback). The key idea for defining a basis gain is derived from spectral decomposition in which the variance of the total decomposition equals the sum of the variances of each individual contribution. Since basis vectors are in general not mutually orthogonal, their contributions do overlap. However, in NMF, coefficients are exclusively non-negative. Consequently, each basis vector b only adds to the total reconstruction of X according to its coefficients C_b and does not delimit other basis vector's contributions. However, it is possible that $C_b b$ explains more variance than what is present in the data. Therefore, the gain of b must be based on how $C_b b$ matches the data X, defined as follows:

$$gain(b) = ||X||_1 - ||X - C_b b||_1 \quad (4)$$

Analysis of this measure facilitates insight into the importance of a basis vector in a factorization. Basis vectors of small gain hint at local minima in the computation, while high gain values, in spite of high errors, suggest at the degree of freedom being set too low for the basis.

4.2. Projection of Factorization Errors

In the following, we describe the design of a visualization that focuses on providing an overview of factorization errors, while highlighting error classes for identifying possible basis vector candidates. Thereby, we rely on two major classifiers for factorization errors:

- **error magnitude** and
- **error irregularity**.

Gaining an overview of a factorization error requires, first and foremost, the visual assessment of the numerical misfit between the original data and its factorization.*Error magnitudes* classify error severeness per data point by a norm. While different norms may be suitable for this task depending on the application, we apply the Euclidean norm to quantify the error magnitudes of SPMS factorization, since it emphasizes larger misfits over smaller ones. Additionally, we classify errors by a measure of *irregularity* that is orthogonal to error magnitudes and suggests a misfit in the factorization basis, as opposed to inadequate numerical computations. This measure of irregularity is defined as follows:

$$\alpha(e) = 1 - \frac{cos\angle(abs(e), \mathbf{1}) - \frac{1}{\sqrt{m}}}{1 - \frac{1}{\sqrt{m}}}, \text{ where} \quad (5)$$

$$cos\angle(abs(e), \mathbf{1}) = ||e||_1 / (||e||_2 \sqrt{m})$$

Thereby, the dominance of a (sparse) feature in the error is defined by the cosine of the angle between its absolute and $\mathbf{1} \in \mathbb{R}^m$, the vector of ones in all coordinates. Independent of the error's magnitude, it holds a measure of irregularity for $0 \leq \alpha(e) \leq 1$, where an error of equal absolute coordinates

leads to a value of 0 and a unit vector to 1. Figure 3 illustrates how this measure is interpreted to SPMS factorization errors. Based on this measure, our projection ϕ, depicting error magnitude and irregularity, is defined as follows:

$$\begin{aligned} \phi : \mathbb{R}^m &\to \mathbb{R}^2 \\ e &\mapsto (\alpha(e), ||e||_2) \end{aligned} \tag{6}$$

Here, e refers to one of n errors, each consisting of m residuals. The y-axis of this projection maps the magnitudes of the factorization errors for each data point, while the x-axis maps to $\alpha(e)$, which measures the dominance of a specific feature in the residuals of an error, as opposed to showing uniform residuals.

When using this mapping, errors of the same magnitude and regularity are mapped to the same locations, regardless of their coordinates being identical. This problem is inherent in dimension reduction and impossible to overcome. However, it can be at least partially alleviated by a color scheme that shows additional differences. The eigenvectors associated with the three largest eigenvalues of the centered error covariance matrix define an orthogonal projection to three dimensions that can assign color values to each error according to its spatial configuration in $\mathbb{R}^m$. Although principal components are less suitable for the mapping of non-orthogonal features than more sophisticated techniques (for example, NMF), it is sufficient for our purpose without burdening interaction time. As such, this color scheme visually differentiates errors that are mapped in proximity while having a different configuration of residuals.

For effective error investigation, the abundance of errors within ranges of specific magnitude and irregularity must be accounted for in the visualization. In order to convey information about the quantity of errors belonging to the same classifiers, the visualization must make aware of the concentration of points within regions of the projection. However, given limited resolution, the specific concentration of points in a projection is visually impossible to assess for large data sets. Although interactively zooming into a projection can unclutter the point configuration, this does not provide quantitative insight into the point concentrations within a region. While assigning opacity values to points, either by the use of alpha blending or by application of a non-linear transfer function, can help convey point density, this approach does not scale well with increasing number of data points.

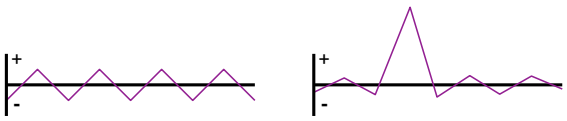


Figure 3: *By utilizing a measure of error regularity (left: regular $\mapsto$ 0, right: irregular $\mapsto$ 1), the presence of dominant features in errors can be quantified, allowing for a visual assessment of noise level.*

In order to convey point concentrations within the projection, we use an approach known as density field contouring. The computation can be summarized as follows. A high-resolution 2D scalar field is computed that holds, for each pixel, the number of points projected to this location. Subsequently, the field is processed via a convolution step using a Gaussian filter kernel, which is scaled to have a peak height of 1 that decreases to 0 over its bandwidth. The Gaussian filter smoothes the field and accumulates density values in the locality of its bandwidth, producing a density field. A texture of contours can be computed, for example, by thresholding for isovalues in the density field. Contours of equal width in image space can be realized by setting the threshold dependent on the local gradient of the density field. For further information on kernel density estimation, we refer to [WJ95]. This method is efficient, and the result is highly effective, depicting quantities of points within regions of the projection.

To summarize the properties of the error visualization defined above, we list the main features in the following:

- **Horizontal axis:** irregularity of errors (feature dominance)
- **Vertical axis:** magnitude of errors (Euclidean norm)
- **Color:** similarity of errors (in $\mathbb{R}^m$)
- **Contours:** local quantity of errors (point density)

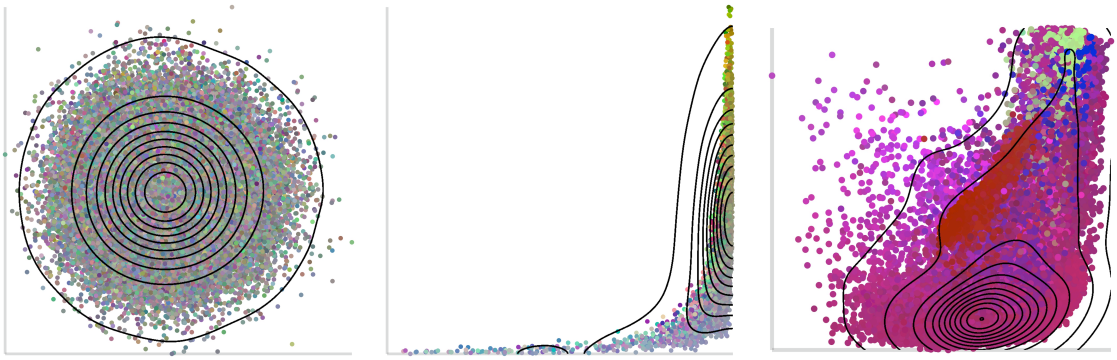
Figure 4 shows examples for different data factorizations.

Interaction

The intuitive visual classifiers described above provide excellent filtering capabilities for high-detail application-specific visualizations. The selection of errors in a specific magnitude-distribution range (regional selection) and/or (sub-)selection of errors based on their spatial relationship (color selection) in this visualization can be linked and act as a filtering mechanism for different high-detail views. Further sub-selection in high-detail views can effectively identify errors that the analyst wants to eliminate. Once errors are identified that are a potential basis vector candidate, we describe in the following how the factorization is updated by incorporation of the candidate into the basis. The scientist can compare the individual gain values and how they would change.

4.3. Interactive Refinement

For facilitating visual comparison and for highlighting changes, we depict the gain values for the basis in a bar chart. After selection of errors, the analysts may request to visualize the potential gain by the addition of a basis vector candidate that eliminates the selected errors with respect to the previous configuration. The optimal basis vector that eliminates select errors is given by the mean of the data points



(a) Errors of Gaussian noise (scaled) (b) Gaussian noise + irregular errors (c) Overview of SPMS factorization errors

Figure 4: *An overview of factorization errors is achieved by projecting errors based on magnitude (vertical axis) and irregularity (horizontal axis). Further classification of error types is provided by color (similarity) and density contours (abundance).*

producing it, weighted by the absolute mean of the errors per coordinate. As such, the basis vector is introduced that has the exact features of the data points that are not covered by the factorization. The coefficient matrix is adjusted projecting all data points onto the candidate vector and adjusting coefficients of the other basis vectors in relation to how the candidate allows for a better representation, while the coefficients for the candidate vector are generated conversely based on the best fit. Using the adjusted starting configuration, NMF is run for several iterations to produce an adequate estimate of the factorization quality that is achievable by including the candidate. Subsequently, the gain of the basis prior to adjustment is visualized in relation to the gain post adjustment in the bar chart, while the differences are highlighted. Figure 6 shows an example to this concept. Interactivity is an integral part of this methodology and performing optimization methods on the GPU is inevitable for large data sets. We describe our implementation briefly in the following.

4.4. Independence Regulation on the GPU

In [WR10], Wilson et al. described a term for regulating mutual independence between the coefficients of basis vectors in non-negative mixtures. Although being very robust, their formulation requires no matrix inversion, making it more flexible than previous approaches and fast to compute on the CPU. The update of the coefficient matrix C, applicable to multiplicative NMF update schemes, that regulates independence is based on the derivative of a cost function J_C measuring correlation coefficients, as described by (3).

We note that the formulation given in [WR10] of the partial derivative $\partial J(C)/\partial C_{a,b}$, is not easily realized on a GPU and can be reformulated more efficiently. By exploiting the fact that the partial derivatives of the correlation matrix terms are symmetric and populated only in a single row and column, we can greatly simplify the formulation as follows:

$$\frac{\partial J(C)}{\partial C_{a,b}} = 4 \left|\left| Corr_{b,\bullet} \otimes \frac{(\mathbf{n}_c\mathbf{n}_c^T)_{b,\bullet} \otimes C_{a,\bullet} - \frac{C_{a,b}}{\mathbf{n}_{c_b}} \mathbf{n}_c \otimes (C^T C)_{b,\bullet}}{\mathbf{n}_c\mathbf{n}_c^{T^2} + \varepsilon} \right|\right|_1 \tag{7}$$

Here, $\otimes$ denotes the element-wise multiplication between two matrices of the same dimensions, analogously to the division of $\mathbf{n}_c\mathbf{n}_c^{T^2}$ which is understood as element-wise division of the element-wise squared outer product matrix of $\mathbf{n}_c$. The correlation matrix $Corr$ and norm vector $\mathbf{n}_c$ are given by

$$\begin{aligned} Corr &= N_C C^T C N_C \,, \\ N_C &= diag(\mathbf{n}_c^{-1}) \,, \text{ and} \\ \mathbf{n}_c &= (||C_{\bullet,1}||_F, ..., ||C_{\bullet,k}||_F) \,. \end{aligned} \tag{8}$$

The formulation (7) requires no index evaluations and only k accumulations for updating each entry in C, as opposed to k^2. Consequently, computations are significantly faster, while being solely based on general operations, lending itself towards a straightforward implementation on the GPU.

5. Results

A case study and domain-expert evaluation by atmospheric scientists describes the value and usability of our method in the following. We have been able to (i) produce factorizations of considerably higher quality 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. Case Study

The data we use as an example was collected from wood stove exhaust using a single particle mass spectrometer

[LR05]. Factorizations of this data are used to quantify emission sources of biomass combustion. This aspect is of interest to atmospheric scientists, as biomass combustion is ubiquitous, while being suspected to play a key role in present day environmental concerns including health effects and climate change. The Pittsburgh June-July data (X) contains roughly 70k particle spectra in 256 dimensions and was factorized (in C and B) using an eight-dimensional basis. The error in this factorization can be quantified in relation to the data, $||X - CB||_F \ / \ ||X||_F$, producing a value of 31.1%. This magnitude of information loss is typical for SPMS factorization, making the need for analysis apparent. In our investigation, we first gain an overview of these errors in the projection shown in the center of Figure 5. The projection quantifies the factorization error per data point based on magnitude (y-axis) and irregularity (x-axis). Examples are shown on the left and right side in the figure. The depth contouring in the projection shows that the majority of the data is factorized with good quality (low error magnitude/irregularity). However, large amounts of spectra are not well approximated. The contours of the projection depict two local maxima in error abundance, reflecting the spectra that are factorized by low and high error magnitude, respectively, while irregularity increases with magnitude.

These results support the initial assumption that there are important features in the data that are not covered by the factorization. Coarse classification of these error classes is provided by the coloring of points in the projection. There are three major error clusters visible in the projection, shown by the local abundance of green, blue, and red points. Selection

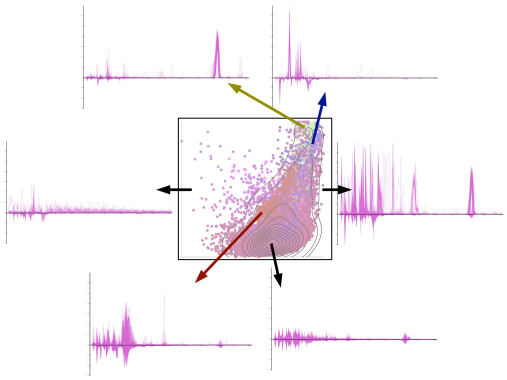
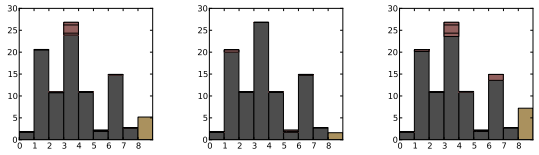


Figure 5: *Errors of the factorization of Pittsburgh source sampling data, June-July, 2002. Selecting errors by color and/or region in the projection (center, also shown in Figure 4(c))) effectively filters high-level views and, thereby, makes possible a detailed data analysis by uncovering errors of high (right) or low (left) irregularity, magnitude, maxima of abundance (bottom right), and provides further classifications by color. Red (bottom left), green (top left), and blue (top right) error clusters are selected.*



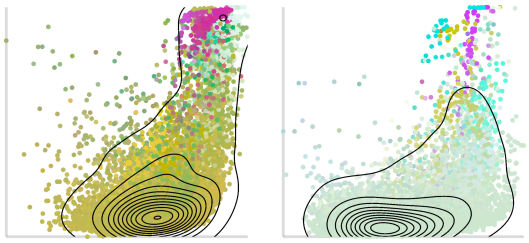
(a) Gain in minimizing green error cluster (b) Gain in minimizing blue error cluster (c) Gain in minimizing red error cluster

Figure 6: *The numerical gain in introducing basis candidates minimizing specific errors is depicted. Sub-optimal parts of the factorization are uncovered by exhibiting a smaller gain than the analysts candidate (left and right). The analyst can add the candidate to the basis, delete existing parts, or continue analysis.*

of these points allows for detailed investigation of the corresponding residuals to be conducted in a high-level view. Such reveals that the error types are characterized by major misfit of the factorization in the following features: (i) $Pb+$-predominant error in green cluster (372 spectra), (ii) NO^+, SiO^+ and Fe^+ in blue cluster (151 spectra), and (iii) $C_xH_y^+$-predominant in red cluster (7,851 spectra).

Having identified dominant error clusters, we investigate the gain in minimizing these errors. Figure 6 shows the estimated improvement that can be gained by introducing a basis vector that minimizes each of the error features. While the (numerical) gain in reducing the error feature outlined by the blue cluster is relatively low, it is considerably higher for the green and red clusters. Noticeably, the gain in introducing a basis vector for these clusters is higher than for other basis vectors (noted by index 0 and 5 in the figure), as computed by the initial factorization. Consequently, we have shown that this basis is sub-optimal and have found alternatives that improve the factorization.

As the initial factorization basis is shown to be sub-optimal in this analysis, the overall error of the factorization can be decreased, while keeping the same dimensionality of the basis. With respect to refining the factorization, the sub-optimal parts of the basis can be deleted and/or the more suitable vectors (for the red and green error classes) added to the basis. Subsequently, the factorization is recomputed with the adjusted basis. In this experiment, we have deleted the sub-optimal parts and introduced the two candidates of higher gain instead. After convergence, the refined factorization features an overall error of 24.7% in relation to the original data. While being restricted to the same dimensionality of the basis as the initial factorization, these results represent an improvement of the overall error by 21.5%. An overview of the remaining error is depicted in Figure 7(a). Noticeably, both error features that were minimized in our refinement are not apparent in the projection. However, there are two new error clusters distinguishable at the top right corner of the



(a) Errors of factorization using an 8-dimensional basis (b) Errors of factorization using a 24-dimensional basis

Figure 7: *By our methodology, atmospheric scientists are able to produce data factorizations of higher quality. An overview is provided of adequate factorization errors with respect to basis dimensionality (8 and 24). (a) By controlled refinement, local minima in the factorization could be uncovered, leading to a decrease of the overall mapping error in relation to the initial solution by 21.5%. (b) Higher quality factorization can be achieved by increasing the dimensionality of the basis, accounting for an overall error of 14.8% in relation to the original (256-dimensional) data.*

projection, in addition to the blue cluster. These new clusters correspond to the two basis vectors that have been deleted in our refinement. Although of high magnitude and irregularity, the clusters contain only a small number of spectra.

Our experiments have shown that significant additional improvement of the factorization for this data set can only be gained by increasing the dimensionality of the basis. However, the amount of information that is consequently added decreases rapidly. Figure 7(b) shows the error projection for a factorization of this data using a 24-dimensional basis. By increasing basis dimensionality, an overall error of 14.8% with respect to the original data was achieved. These results make apparent the need for visual analysis in data factorization. Looking beyond the scope of this work, results also indicate that more research needs to be conducted to support application domains. As such, actively searching for specific error features may provide analysts with the ability to query factorization errors and to quantify the quality of the approximation with respect to these features.

5.2. Expert Evaluation

The recent advent of single particle and related real time techniques in atmospheric science has increased the quality and quantity of available data, so that improvements in data visualization and comprehension techniques are increasingly desired. Single particle mass spectrometers and other similar instruments that collect spectra in real time generate a tremendous amount of data of high dimensionality. These huge, complex data sets pose challenges for atmospheric scientists that need to analyze the data for various endpoints such as emissions source, atmospheric transformations and toxicity. The high dimensionality of the data also confounds comprehension by the atmospheric scientist because so few dimensions can be readily observed.

The methods presented here reduce the dimension of the data set dramatically by discovering the bases that underlie the data and visually present the resulting information to the scientist in a way that elucidates the factors that establish the basis as representing significant pollutant sources or atmospheric transformations. In typical studies, the common bases are hundreds or thousands of times more prevalent than the uncommon ones so techniques for identifying the bases must also take into account that bases with infrequent spectra may have lower variability so appear more significant. Data analysis must not arbitrarily exclude this important information but instead communicate important basis properties, such as efficiency, local minima, and information loss, to the scientist.

Our system supports this objective and enables more accurate and verifiable data analysis. The visualization makes it possible to analyze and classify different basis sets with respect to information loss and different objectives. Alternative basis configurations can be readily identified, by a cluster in the projection, and then selected for analysis. Visually comparing the efficiency of basis vectors enables one to explore alternatives and identify new bases. The interactive nature of this new tool enables ready exploration of hypotheses and discovery of aspects of such large data sets that one might not be able to discover otherwise.

6. Conclusions

It is very important and difficult to address the issue of "error" in any data factorization method and application setting. In our case, error can be inherently associated with the result of approximating original data in a lower-dimensional space. Error magnitude and meaning are directly influenced by the number of chosen basis vectors and the efficiency of the basis transformation. This multi-criteria and non-convex optimization problem cannot be solved in an optimal way by known algorithms. It is therefore crucially important to have the data analyst play an integral role in the entire process of factorization: it is the expert's insight and understanding of a specific problem - understanding a set of several thousand 256-dimensional data in our case - that allows us to have a user involved in specifying the number of dimensions needed for lower-dimensional approximation, in specifying individual basis vectors, and in determining what is and what is not a "good approximation." Error quantification and visualization, combined with the ability to interactively influence the data factorization/approximation process, is thus a necessary component of any system aimed at dramatically reducing the dimensionality of a complex and high-dimensional data set to assist effectively with understanding. Our approach is exactly supporting this objective.

References

[CJ10] COMON P., JUTTEN C.: *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Academic Press, 2010. 2

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics 14* (2008), 1141–1148. 2

[EGG*12] ENGEL D., GREFF K., GARTH C., BEIN K., WEXLER A. S., HAMANN B., HAGEN H.: Visual steering and verification of mass spectrometry data factorization in air quality research. *IEEE Trans. Vis. Comput. Graph. 18*, 12 (2012), 2275–2284. 2

[EMO*11] ESSER E., MÖLLER M., OSHER S., SAPIRO G., XIN J.: A convex model for non-negative matrix factorization and dimensionality reduction on physical space. *Arxiv preprint arXiv11020844 stat.ML* (2011), 14. 2

[FR11] FERDOSI B. J., ROERDINK J. B. T. M.: Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. *Comput. Graph. Forum 30*, 3 (2011), 1121–1130. 2

[GTC01] GRINSTEIN G., TRUTSCHL M., CVEK U.: High-dimensional visualizations. In *Proceedings of Visual Data Mining workshop, KDD'2001* (2001). 2

[HLD02] HAUSER H., LEDERMANN F., DOLEISCH H.: Angular brushing of extended parallel coordinates. In *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)* (Washington, DC, USA, 2002), IEEE Computer Society, p. 127. 2

[IMO09] INGRAM S., MUNZNER T., OLANO M.: Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 249–261. 2

[Ins09] INSELBERG A.: *Parallel Coordinates*. Springer, 2009. 2

[JBS08] JÄNICKE H., BÖTTINGER M., SCHEUERMANN G.: Brushing of attribute clouds for the visualization of multivariate data. *IEEE Transactions on Visualization and Computer Graphics 14* (2008), 1459–1466. 2

[JLJC05] JOHANSSON J., LJUNG P., JERN M., COOPER M.: Revealing structure within clustered parallel coordinates displays. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 17–. 2

[KBHH05] KIM E., BROWN S. G., HAFNER H. R., HOPKE P. K.: Characterization of non-methane volatile organic compounds sources in houston during 2001 using positive matrix factorization. *Atmospheric Environment 39*, 32 (2005), 5934–5946. 2

[KP08] KIM J., PARK H.: Fast nonnegative matrix factorization: an active-set-like method and comparisons. *Science* (2008). 2, 3

[LGD*05] LARAMEE R. S., GARTH C., DOLEISCH H., SCHNEIDER J., HAUSER H., HAGEN H.: Visual analysis and exploration of fluid flow in a cooling jacket. In *In Proceedings IEEE Visualization 2005* (2005), pp. 623–630. 2

[LR05] LIPSKY E., ROBINSON A.: Design and evaluation of a portable dilution sampling system for measuring fine particle emissions from combustion systems. *Aerosol Science and Technology 39*, 6 (2005), 542–553. 8

[LS00] LEE D. D., SEUNG H. S.: Algorithms for non-negative matrix factorization. In *In NIPS* (2000), MIT Press, pp. 556–562. 3

[OHJS10] OESTERLING P., HEINE C., JÄNICKE H., SCHEUERMANN G.: Visual analysis of high dimensional point clouds using topological landscapes. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE* (Mar. 2010), pp. 113 –120. 2

[PEP*11] PAULOVICH F., ELER D., POCO J., BOTHA C., MINGHIM R., NONATO L.: Piece wise laplacian-based projection for interactive data exploration and organization. *Computer Graphics Forum 30*, 3 (2011), 1091–1100. 2

[RZH12] ROSENBAUM R., ZHI J., HAMANN B.: Progressive parallel coordinates. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE* (28 2012-march 2 2012), pp. 25 –32. 2

[SLY*09] STUMP G., LEGO S., YUKISH M., SIMPSON T. W., DONNDELINGER J. A.: Visual steering commands for trade space exploration: User-guided sampling with example. *Journal of Computing and Information Science in Engineering 9*, 4 (2009), 044501. 2

[WBP07] WEBER G., BREMER P.-T., PASCUCCI V.: Topological landscapes: A terrain metaphor for scientific data. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (Nov.-Dec. 2007), 1416 –1423. 2

[WFR*10] WASER J., FUCHS R., RIBICIC H., SCHINDLER B., BLÖSCHL G., GRÖLLER E.: World lines. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (nov.-dec. 2010), 1458 –1467. 2

[WGK10] WARD M. O., GRINSTEIN G., KEIM D. A.: *Interactive Data Visualization: Foundations, Techniques, and Application*. A. K. Peters, Ltd, 2010. 2

[WJ95] WAND M. P., JONES M. C.: *Kernel Smoothing*, vol. 60. Chapman & Hall/CRC, 1995. 6

[WR10] WILSON K. W., RAJ B.: Spectrogram dimensionality reduction with independence constraints. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* (march 2010), pp. 1938 –1941. 2, 7

[YPWR03] YANG J., PENG W., WARD M. O., RUNDENSTEINER E. A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symposium on Information Visualization* (2003). 2

[ZIN*08] ZELENYUK A., IMRE D., NAM E. J., HAN Y., MUELLER K.: Clustersculptor: Software for expert-steered classification of single particle mass spectra. *International Journal of Mass Spectrometry 275*, 1âĂŞ3 (2008), 1 – 10. 2