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

D. Engel[1], M. Hummel[1], F. Hoepel[1], K. Bein[2], A. Wexler[2], C. Garth[1], B. Hamann[3], and H. Hagen[1]

[1]University of Kaiserslautern, Germany; [2]Air Quality Research Center (AQRC), University of California, Davis, CA, USA;
[3]Institute for Data Analysis and Visualization (IDAV), Department of Computer Science, University of California, Davis, CA, USA

Abstract

Analysis of chemical constituents from mass spectrometry of aerosols involves non-negative matrix factorization, an approximation of high-dimensional data in lower-dimensional space. The associated optimization problem is non-convex, resulting in crude approximation errors that are not accessible to scientists. To address this shortcoming, we introduce a new methodology for user-guided error-aware data factorization that entails an assessment of the amount of information contributed by each dimension of the approximation, an effective combination of visualization techniques to highlight, filter, and analyze error features, as well as a novel means to interactively refine factorizations. A case study and the domain-expert feedback provided by the collaborating atmospheric scientists illustrate that our method effectively communicates errors of such numerical optimization results and facilitates the computation of high-quality data factorizations in a simple and intuitive manner.

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. An important step in tackling this problem is to elucidate the chemical compounds of ambient airborne particles. A single particle mass spectrometer (SPMS) now chemically analyzes individual aerosol particles in real time, providing unprecedentedly rich data sets for air quality and climate research. These data sets contain the mass spectra of collected particles, thereby describing particles based on their distribution of ions by mass. An exemplary mass spectrum is shown in Figure 1. These histograms are stored and interpreted as points in high-dimensional space for consecutive analysis of their comprised chemical compounds.

As mass is ambiguous, various ions may contribute to each coordinate/dimension and SPMS mass spectra do not lend themselves to a straightforward deduction of chemical compounds. The corresponding optimization problem can be described as follows: given data that is derived from a combination of unknown compounds in unknown abundance and combination, the goal is to factor out both unknowns, provided only with an estimate of the number of compounds and an assumption of their mixing model. A basis transformation is to be found that models the observed spectra as a linear combination of the spectra arising from each of the chemical compounds, such that linear combinations of this basis forms the observed mass spectra. Physical and chemical constraints further dictate both basis and coefficients to be non-negative, rendering spectral decompositions inapplicable and the optimization problem non-convex. Consequently, approximations can only be computed iteratively by gradient-based approaches such as non-negative matrix factorization (NMF) [EGG*12] and currently known algorithms can only produce sub-optimal results. The approximation error, defined as the discrepancy between data and its lower-dimensional approximation, of mere locally optimal approximations deviates significantly from that of a globally optimal solution. Therefore, the quality of approximations needs to be assessed by the scientist. The visual communication of errors in non-negative matrix factorization, however, has not yet been studied in visualization research and common visualization tools are not applicable to this problem.

We introduce a new methodology to visual analysis of approximation errors in non-negative matrix factorization that includes (i) an approach to assess the quality of a factorization basis, (ii) a visualization of factorization errors, and

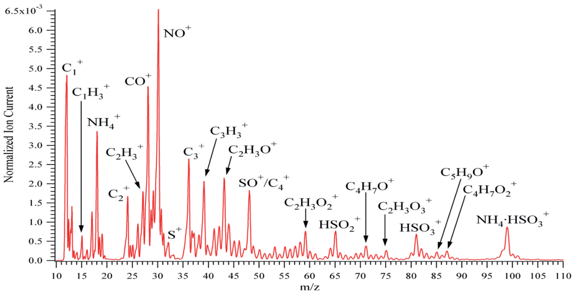


Figure 1: *The mass spectrum of an aerosol represents a pattern (coordinates) that quantifies the abundance of inherent fragment ions (peak labels) per mass (dimensions). Data factorization provides lower-dimensional representations of aerosols in terms of latent components of these patterns.*

(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 improve the factorization quality and consequently overcome "being stuck" in local minima of non-convex optimization. Due to the high degree of interactivity 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 regarding the degree of freedom needed to approximate high-dimensional data. Thereby, we contribute to applications including, but not limited to, air quality research, by providing novel means to elucidate chemical species content of MS spectra from any wide range of sources. The core of our methodology can further be generalized and applied to other settings of non-convex linear approximation of high-dimensional data.

The remainder of the paper is structured as follows. Section 2 discusses related work in matrix 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 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

Non-negative matrix factorization (NMF) [CJ10] computes a non-negative linear basis transformation that approximates high-dimensional data in lower-dimensional space. Thereby, a matrix is factorized into two matrices, representing a basis and corresponding coefficients. In air quality research, NMF is used to identify chemical constituents and consequently classify spectra in particle types [KBHH05]. 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 spectra in terms of non-negative combination of latent physical components. Mass is inherently non-negative, as is the composition of spectra into components, 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. In addition to finding an optimal solution, interpretability is often the greatest problem when working with dimension reduction. Making results 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]. 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]), uncertainty visualization [CCM09], 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 [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. Being based on parallel coordinates [Ins09], the visualization can become highly dense

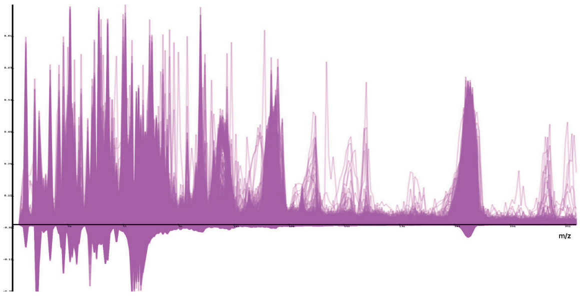


Figure 2: *Previous work visualizes the errors produced by SPMS data factorization in high detail. Due to data complexity and dimensionality, this representation is prone to visual clutter and fails to provide an overview to analysts who are faced with the problem of identifying, classifying, and analyzing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

and cluttered, rendering it unsuitable to analyze factorization errors for large data sets without overview. An example is given in Figure 2. In contrast, the present work focuses on visualizing and analyzing these errors. We introduce a new 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.

3. Requirements Analysis

In the following, a brief account of the application background is given, as well as 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 to analyze individual aerosol particles at sampling sites of atmospheric interest. Processing of these mass spectra enables atmospheric scientists to categorize the large number of spectra obtained into useful clusters of identified particle types. The mass spectrum of a particle represents a function that maps the abundance of fragment ions per *mass over elemental charge (m/z)*. Discretized in bins of 1 m/z, the SPMS analyzer employed in this work captures the first 256 m/z ratios. The resulting histogram is stored as a 256-dimensional vector, where each coordinate corresponds to the abundance of fragment ions 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 - the particle's chemical constituents. 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 $X = CB + N$. $B \in \mathbb{R}_+{}^{(k \times m)}$ is the matrix storing (row-wise) basis vectors, such that X is derived with the coefficient matrix $C \in \mathbb{R}_+{}^{(n \times k)}$ 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, NMF can cope with these conditions and produce viable solutions to the problem.

Non-negative matrix factorization (NMF) computes a basis B and coefficients C, by minimizing the mapping error,

$$J = ||X - CB||_F^2 \ \rightarrow \ \min, \tag{1}$$

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 an alternating 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 and we first update C while keeping B fix. Thus, if the basis is initially globally optimal, the optimization converges to equally optimal coefficients.

In addition to minimizing the overall mapping error J, feature independence is imposed to the optimization. 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 and defined by the squared Frobenius norm of the uncentered correlation matrix,

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

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 a slow convergence speed, a proneness to become "stuck" in local optima, and the required input on the number of basis vectors k. 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 trade-off between basis dimensionality and approximation error.

3.2. Errors in SPMS data factorization

Several 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 approximation to lower-dimensional space produces errors. However, given the non-convex nature of our optimization, for which globally optimal results cannot be expected, analyzing these errors becomes a necessity. 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. 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 are those that significantly deviate from a Gaussian distribution. If these specific error features occur in abundance, it indicates 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. 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 due to numerical inefficiency. To determine what may serve as adequate to the current purpose of analysis is left to the analyst.

3.3. Tasks and Requirements

In order to determine an adequate factorization of SPMS data, atmospheric scientists have the ultimate task to minimize both dimensionality and error of the approximation. Thereby, the goal is to choose a trade-off between dimensionality and error, admitting identified errors that have been minimized and are unavoidable due to dimension reduction for the sake of having a lower-dimensional representation. However, for mere locally optimal solutions, it is unclear whether errors are truly minimized and unavoidable. Therefore, a methodology is needed to assess both (i) the error and (ii) the quality of a factorization. While the overall approximation error can be computed as described in the previous section, the quality of a factorization relates to the efficiency of a basis in approximating the data. Basis efficiency quantifies how much information from the data is represented in the factorization in relation to a globally optimal solution given the same degree of freedom. As knowledge of a globally optimal solution is unknown in general, analyzing factorization quality requires a human-in-the-loop approach and the tools to aid in visual analysis.

It is only by the conveyance of both properties (error and efficiency) that scientists can determine the “right” dimensionality for the basis and an adequate approximation of the data. Analysis to ascertain basis efficiency must be tightly coupled with the visualization of error features (and their significance) to aid the scientist in determining an admissible trade-off, deciding which errors to admit as a consequence of dimension reduction and weighting errors against dimensionality of the approximation. 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. **Visualizing error features:**
   Error visualization should convey a classification of errors by importance and type, and serve as a basis to conduct detailed analysis. A major requirement is the visual separation of noise from specific error features described in the previous section. The visualization must convey 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, as described in Section 3.2.
2. **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, when possible, uncover inefficiencies of the factorization basis with respect to approximating the data.
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 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

We describe how factorization quality can be analyzed, suboptimality assessed and the factorization be improved. Essential to our approach is a highly visual and analytical framework that involves the analyst in several key steps.

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 basis must hold that there are better basis vectors that produce less overall error, with respect to lowering error magnitudes in their abundance (accumulation according to (1)). Conversely, the presence of similar errors of high magnitude and abundance directly corresponds to a basis vector candidate that is not part of the current basis, while being numerically beneficial to be included. This leads one to conclude that optimality of the basis can be assessed by comparing the amount of information currently conveyed by each basis vector against the amount of information that could be conveyed by basis vector candidates. Further, these candidates are directly reflected by and can be identified based on similar errors of high magnitude and abundance. Consequently, local minima in the factorization can be revealed by (i) identifying candidates based on visually assessing error magnitudes, similarity, and abundance, and (ii) comparing the amount of information conveyed by the current basis vectors in relation to that of the candidate. If the candidate allows for the conveyance of more information than one of the current basis vectors, then the basis is sub-optimal. In this case, the candidate can be introduced into the basis, possibly by exchanging one of the other basis vectors of lower benefit.This concept requires three aspects: (i) a comparative measure of conveyed information per basis vector, (ii) an error visualization focusing on error magnitudes, similarity, and abundance, and (iii) interactive probing of the error to visually compare the benefit of each basis vector against that of selected candidates. We introduce this measure in the following.

In NMF, coefficients are exclusively non-negative. Consequently, each basis vector $b_i \in \mathbb{R}^m$ only adds to the total approximation of $X \in \mathbb{R}^{(n \times m)}$ according to its coefficients $c_i \in \mathbb{R}^n$ and does not delimit other basis vectors' contributions. Thus, the overall approximation is decomposed into each basis vector's contribution, such that $CB = \sum_{1 \leq i \leq k} c_i \otimes b_i$. For each basis vector's contribution, we quantify its "gain" by a norm of the residuals to X. It is possible that such contributions cover more variance than present in X. Therefore, the gain of b_i must be based on how its contribution matches the data. Consequently, our gain measure is defined as follows:

$$gain(b_i) = ||X||_1 - ||X - c_i \otimes b_i||_1 \; . \qquad (3)$$

Analysis of this measure facilitates insight into the amount of information each basis vector introduces in the factorization. Thus, spotting a local optimum reduces to identifying the basis vectors of small gain and comparing them to the gain of the basis vector candidate that corresponds to the largest error cluster.

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: **magnitude** and **irregularity**. *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*, similar to Hoyer's sparsity measure [HR09], that is orthogonal to error magnitudes and suggests a misfit in the factorization basis, as opposed to inadequate numerical computations. With $e \in \mathbb{R}^m$ referring to one of n errors, each consisting of m residuals, this measure of error irregularity is defined as follows:

$$\begin{aligned} \alpha(e) &= 1 - \frac{cos\angle(abs(e), \mathbf{1}) - \frac{1}{\sqrt{m}}}{1 - \frac{1}{\sqrt{m}}}, \text{ where} \qquad (4) \\ cos\angle(abs(e), \mathbf{1}) &= ||e||_1 / (||e||_2 \sqrt{m}) \; . \end{aligned}$$

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

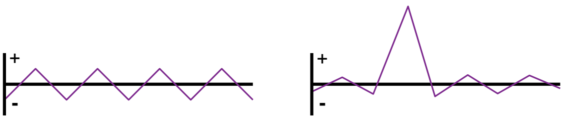
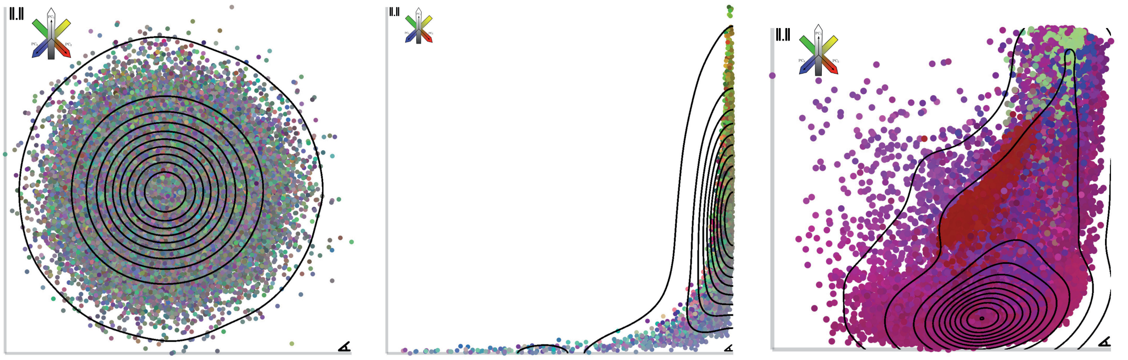


Figure 3: *By utilizing a measure of error regularity (left: regular ↦ 0, right: irregular ↦ 1), the presence of dominant features in errors can be quantified, allowing for a visual assessment of noise level.*



(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).*

error magnitude and irregularity, is defined as follows:

$$\begin{aligned} \phi: \ \mathbb{R}^m &\to \mathbb{R}^2 \\ e &\mapsto (\alpha(e),\ ||e||_2) \end{aligned} \qquad (5)$$

The y-axis of this projection maps the magnitudes of the factorization errors, while the x-axis maps to $\alpha(e)$, which indicates the dominance of a specific feature in the residuals of an error, as opposed to showing uniform residuals.

By 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 to dimension reduction and impossible to overcome. However, it can be at least partially alleviated by a color scheme that shows additional differences. We define a projection to three dimensions that assigns color values to each error according to its spatial configuration in $\mathbb{R}^m$. First, the errors $X - CB$ are normalized to unit scale in order to render the projection independently of error magnitudes and centered in order to utilize the full color range. Second, the covariance matrix is built from this normalized centered matrix. Finally, the eigenvectors associated with the three largest eigenvalues of this covariance matrix define the projection into color space. A suitable color space is, for example, CIE*lab*, as it is uniform and of orthogonal basis.

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.

In order to convey point concentrations within the projection, we use density field contouring. Thereby, 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].

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 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 error features of a factorization. These features correspond to a potential basis vector candidate that eliminates these errors. In the following, we describe how the factorization quality can be assessed based on these candidates and how the basis can be interactively refined.

4.3. Interactive Refinement

After the selection of errors that are of interest in the interactive analysis process, our methodology entails the visualization of the potential gain produced by the addition of the corresponding basis vector candidate. This candidate, the optimal basis vector that eliminates the selected errors, is given by the mean of the data points 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 this adjusted starting configuration, our NMF model, as described in Section 3.1, is performed for several iterations to produce an adequate estimate of the factorization quality that is achievable by including the candidate. Due to the linear nature of the approximation, the gain that can be expected by the addition of a vector to the basis depends on the magnitude and abundance of spectra covered by the vector minus the variance between the spectra. Consequently, the gain is highest for introduction of a basis vector corresponding to an abundance of large errors showing similar features. However, by defining a basis candidate, the analyst does not restrict the basis. While the two-block optimization scheme will first optimize the coefficients to the initial basis, consecutive iterations will also update the basis vectors if they are not optimal. Without delimiting the optimization, this methodology can be used to overcome local minima, as well as to analyze and refine the basis.

Subsequently to the NMF optimization with adjusted basis, the gain of the basis prior to adjustment is visualized in relation to the gain post adjustment in the chart, while the differences are highlighted in distinct colors. Figure 6 shows an example of this separate view in our framework. If the gain of the analysts candidate is larger than that of a basis vector from the previous basis configuration, then this candidate contributes more information to the approximation and, consequently, a local minimum in the computation has been uncovered. On the other hand, equally high gain values for all basis vectors, in spite of high errors, suggest that the degree of freedom is set too low for the basis. By selection in the bar chart, the analyst can flag any basis vector to be added or removed from the basis and subsequently trigger the optimization to be performed again for the desired configuration. Thus, the basis vector that minimizes the selected errors can be added to the basis, other basis vectors of low gain can be deleted, or the candidate can be forfeited in order to continue probing of the errors. As interactivity is an integral part of this analysis, performing optimization methods on the GPU is inevitable for large data sets. We describe our implementation 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 (2).

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 non-zero only in a single row and column, we can greatly simplify the formulation as follows:

$$\frac{\partial J(C)}{\partial C_{a,b}} = 4 \left|\left| Corr_{b,\bullet} \otimes \frac{(\mathbf{n}_c \mathbf{n}_c^T)_{b,\bullet} \otimes C_{a,\bullet} - \frac{C_{a,b}}{\mathbf{n}_{c_b}} \mathbf{n}_c \otimes (C^T C)_{b,\bullet}}{\mathbf{n}_c \mathbf{n}_c^{T^2} + \varepsilon} \right|\right|_1 \tag{6}$$

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

$$\begin{aligned} Corr &= N_C C^T C N_C \,, \\ N_C &= diag(\mathbf{n}_c^{-1}) \,, \text{ and} \\ \mathbf{n}_c &= (||C_{\bullet,1}||_F, \ldots, ||C_{\bullet,k}||_F) \,. \end{aligned} \tag{7}$$

The formulation (6) 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

The following case study and domain-expert feedback provided by atmospheric scientists demonstrates the utility of our method. 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 based on error magnitude (y-axis) and irregularity (x-axis). Snapshots from a detailed view of selected errors are shown on the sides 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 of these points allows for detailed investigation of the corresponding residuals to be conducted in a high-level view. This 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).

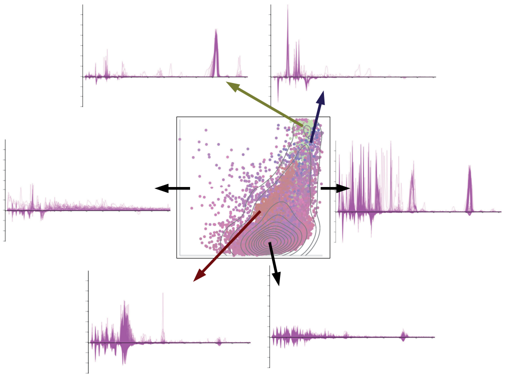
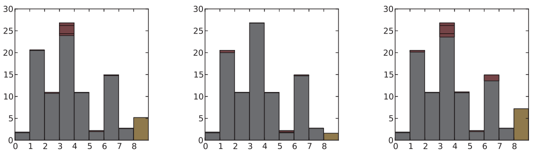


Figure 5: *Errors of the factorization of Pittsburgh source sampling data, June-July, 2002. Selecting errors by color and/or region in the projection (center, also shown in Figure 4(c)) effectively filters high-level views and, thereby, makes possible a detailed data analysis by uncovering errors of 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.*

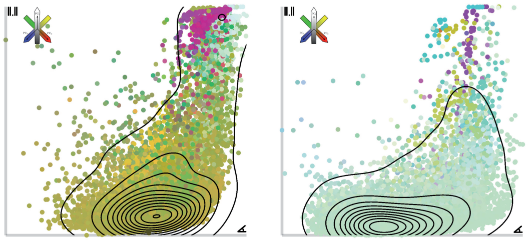


(a) Gain in minimizing green errors (b) Gain in minimizing blue errors (c) Gain in minimizing red errors

Figure 6: *The numerical gain in introducing basis candidates minimizing specific errors is depicted in relation to the previous basis configuration (red = decrease). Sub-optimal parts of the factorization exhibit a smaller gain than the analysts candidate ((a) and (c)). The analyst can add the candidate to the basis, delete existing parts, or continue analysis.*

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 suboptimal 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 suboptimal 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



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

Figure 7: *(a) Controlled refinement of the factorization produced a decrease of the overall error by 21.5% in relation to the initial solution. (b) Further decrease was achieved by increasing the basis dimensionality, here accounting for an overall error of 14.8% in relation to the original data.*

not apparent in the projection. However, there are two new error clusters distinguishable at the top right corner of the 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 Feedback

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 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. The system described here 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, ultimately producing factorizations of higher quality. 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 important and difficult to address the issue of “error” in any data factorization method and application setting. In our case, error can be associated with the result of approximating original data in a lower-dimensional space. The error is 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: by specifying the number of dimensions needed for lower-dimensional approximation, specifying individual basis vectors, and 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 highly desirable and crucially important 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.

Acknowledgements

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

References

[CCM09] CORREA C., CHAN Y.-H., MA K.-L.: A framework for uncertainty-aware visual analytics. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2009), pp. 51–58. 2

[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. 1, 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

[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)* (2002), pp. 127–130. 2

[HR09] HURLEY N., RICKARD S.: Comparing measures of sparsity. *IEEE Transactions on Information Theory 55*, 10 (2009), 4723–4741. 5

[IMO09] INGRAM S., MUNZNER T., OLANO M.: Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics 15*, 2 (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*, 6 (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 *IEEE Pacific Visualization Symposium (PacificVis)* (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. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (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 *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (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*, 1-3 (2008), 1–10. 2