Response to July 2023 Review

Sean Youn

July 2024

This document addresses comments from reviewers for the article "Positive Matrix Factorization of Large Real-Time Atmospheric Mass Spectrometry Datasets Using Error-Weighted Randomized Hierarchical Alternating Least Squares" (Sapper et al.). Note the change in the title (previously "Positive Matrix Factorization of Large Aerosol Mass Spectrometry Datasets Using Error-Weighted Randomized Hierarchical Alternating Least Squares") in order to be more precise with regards to the principal dataset used in the presented analysis.

The original comments by the reviewer are presented in red text, our responses are shown in black text, and significant additions to the manuscript are presented in blue text. In the time since the last submission of this manuscript, we have added two new co-authors (Youn and Jimenez). Some additional changes only affecting readability of the manuscript have been made as well.

1 Reviewer 1

No comments or criticisms to address

2 Reviewer 2

2.1 Paragraph 1

In the Appendix there was a statement that H is a kxn matrix that is assumed to be full column rank (even though k was assumed to be smaller than n). In their response, the authors say that this (erroneous) statement was corrected with "H is a kxn matrix that is assumed to be full row rank". However, the revised document (line 666) still has the wrong statement.

The text in question is present in line 630 in the updated manuscript and has been modified appropriately.

However, there seem to be cases where this [Hadamard notation] was not applied, e.g. lines 158 and 161 (there might be others)

Equations 1, 3, 11, 19, 20, 21, 22, 23, 24, 38, and A1 and lines 75, 89, 259, 293, and 405 have been updated to incorporate Hadamard notation for element-wise division in particular. Note that line numbers have changed between the version

viewed by the reviewer and the current manuscript due to other changes in the main text. We believe all presented equations now make full use of Hadamard notation when appropriate as suggested by the reviewer.

2.2 Paragraph 2

The authors claim in the abstract that their algorithm results in computational speedups of 38, 67 and 634 compared to other algorithms. The statement is not justified by the methodology adopted in this paper. In particular, there is no detailed performance evaluation, neither a careful optimization of the methods. A modest statement, e.g. "numerical experiments with the proposed method indicate that the method is faster than competing algorithms" would be more appropriate

In order to address this comment and other comments in 2.3, the abstract text will be changed to the following:

"Weighted positive matrix factorization (PMF) has been used by scientists to find small sets of underlying factors in environmental data. However, as the size of the data has grown, increasing computational costs have made it impractical to use traditional methods for this factorization. In this paper, we present a new external weighting method to dramatically decrease computational costs for these traditional algorithms. The external weighting scheme, along with the Randomized Hierarchical Alternating Least Squares (RHALS) algorithm, was applied to the Southern Oxidant and Aerosol Study (SOAS 2013) dataset of gaseous highly oxidized multifunctional molecules (HOMs). The modified RHALS algorithm successfully reproduced six previously-identified, interpretable factors with the total computation time of the non-optimized code showing potential improvements on the order of one to two orders of magnitude compared to competing algorithms. We also investigate rotational ambiguity in the solution, and present a simple "pulling" method to rotate a set of factors. This method is shown to find alternative solutions, and in some cases, lower the weighted residual error of the algorithm."

2.3 Paragraph 3

This paper essentially uses a modification of the NMF to target a specific dataset (analyzed extensively by Massoli et al.) for which k=6 factors have been found to be appropriate. The paper makes no effort to consider different combinations of (m,n,k). It seems to me that this makes the paper very application and data specific and possibly not as interesting. In any case, the authors should declare this early in the paper so that the readers are not misled regarding the generality of the method.

Text and wording to clarify the main application/study of the algorithm to SOAS (2013) have been included in the revised abstract already shown in 2.2. However, experimentation on matrices of different sizes, the predominant variable between real-time datasets obtained with different mass spectrometers, was done in this study (see Section 4.2 (Simple Case) and Figure 4) and is

discussed accordingly. Analysis of the operations per step for each algorithm is also presented in Table 1 for general values of m, n, and k. In addition, five or six factors are typical in real-time atmospheric and aerosol mass spectrometry data (Ulbrich et al. (2009); Massoli et al. (2018); Zhang et al. (2011)). Though variation in the optimal number of factors can theoretically occur depending on the dataset, significant deviation from five or six factors is rare in practice.

Real-time mass spectrometry data sets are variable in terms of the matrix size, and we believe this constitutes enough diversity to be useful to the general atmospheric/aerosol mass spectrometry community. To reflect some of the reviewer's concerns, the following addition will be made to Section 1.7 (Data):

"We use this six factor solution as a reference solution, and test whether the RHALS algorithm can recreate formulated factors as well as those found from PMF2. Analysis of results for different numbers of factors (other than the original six identified in Massoli et al. (2018)) were not considered in order to maintain interpretability of the algorithm output. For reference, the PMF2 factor mass spectra and the time trends over all of the data are shown in Figure 1. The factor time series, as well as the time series of the total mass concentration is also shown in Figure 2. Both plots show total concentration amounts over the entire time series and mass spectra respectively."

2.4 Paragraph 4

The authors insist on using the term PMF instead of NMF (because of the use of the term by Paatero et al. However, the matrices used in the paper are not positive, they are nonnegative. Calling a matrix with some zero elements positive is mathematically incorrect, with all due respect to the pioneering work of Paatero et al. Indeed, there are results that hold for positive matrices but do not hold for nonnegative ones unless extra conditions are imposed (e.g. the Perron theorem

We acknowledge that the reviewer is entirely correct in their assertion that nonnegative matrix factorization is a more precise term than positive matrix factorization from a mathematical perspective (and that the two are not necessarily always equivalent). However, positive matrix factorization remains the dominant nomenclature, especially in the community most relevant to this study due to the explicit use of the term PMF by the EPA for their model/software (by Paatero). As of this writing, a quick search in Google Scholar lists about 506,000 results for "positive matrix factorization" and about 202,000 results for "nonnegative matrix factorization." Furthermore, a search of "positive matrix factorization aerosol" yields 16,800 results compared to 5,170 from "nonnegative matrix factorization aerosol."

Nevertheless, to avoid any further confusion, we have made the following modification to line 36 in Section 1.1 (Problem Statement):

"In Eq.(1), $\oslash$ represents elementwise division, the norm $||\cdot||_F$ is the Frobenius norm, and all elements of $\mathbf{W}$ and $\mathbf{H}$ are constrained to be nonnegative. Further, we note that for consistency with nomenclature in the literature related to use of this algorithm for factorization of aerosol mass spectrometry datasets, we refer

to this approach as "positive" matrix factorization (i.e., PMF) while recognizing that a more precise name would be nonnegative matrix factorization."

2.5 Paragraph 5

The statement in the authors' response "We have addressed the Computational Cost in 2 in response to Main Criticism 1. Both proposed postprocessing steps are O(mnk), and we found experimentally that using MATLAB's pinv, a Krylov subspace method (Feng et al.,2018), generated slightly faster results." is wrong, at least regarding the MATLAB pinv function. It is likely that the authors were misled by the discussion in the reference by Feng et al. which argues (rightly) for using the lansvd of PROPACK instead of the Mathworks svds for performing large scale truncated svd. On the other hand, the native MATLAB pinv does not use svds but the native MATLAB svd function which is certainly not a Krylov method. Incidentally, this is another example where the choice of citation matters. This affects lines 305-311 of the text and possibly others.

The section in question is in Section 2.3 (External Weighting). In order to avoid confusion regarding the issue raised by the reviewer, reference to MATLAB's pinv function as a Krylov subspace method was removed from the text as it is not critical to understanding the main mechanisms of the presented algorithms.

2.6 Paragraph 6

Throughout the paper, the authors use the symbol k sometimes as index and other times as the rank of the sought factorization which is confusing. Also, in some summation formulas (e.g. 2), the authors use summation (e.g. over k) without specifying the range while in others (e.g. 7) they do.

Notation has been standardized to refer to m, n, and k as the number of rows, columns, and factors respectively and to to use i, j, and l to refer to indices of the rows, columns, and factors respectively (as is needed in summation notation, among other cases). Summations have also been modified to explicitly state the ranges of summation in all cases.

If Q_j is defined as stated (minimizing) then the min operator on the left-hand side of relation (11) is redundant.

The true cost function is defined as follows:

$$Q_j = ||(\mathbf{R}_j - \mathbf{W}(:,j)\mathbf{H}(j,:)) \oslash \mathbf{\Sigma}||_F^2 \tag{1}$$

Minimizing the cost-function is purpose of the algorithm, but the cost function itself is not the minimization of the equation as currently presented. References to the cost function have been modified to match the equation in 1 and the wording/equation in Section 2.1 (HALS Algorithm)) has been modified to the following:

"The HALS algorithm applies block coordinate descent methods in order to minimize the cost function Q_j by minimizing a "block," or outer product of

individual factors, of $\mathbf{W}$ and $\mathbf{H}$ at a time while keeping the other factors fixed. (Erichson et al. (2018)).”

$$Q_j = ||(\mathbf{R}_j - \mathbf{W}_{(:,j)}\mathbf{H}_{(j,:)}) \oslash \mathbf{\Sigma}||_F^2 \tag{2}$$

Using both Q_j^i and Q_j^P (defined differently) is confusing.

Q_j^i and Q_j^p are first introduced and defined in Equations 13 and 14 respectively. Lines 228 to 230 (immediately preceding these equations) are modified to the following to make clear we are differentiating between rows and columns of the matrices:

“To derive update rules for HALS, partial derivatives of Eq. (1) are taken with respect to the factors $\mathbf{W}_{(:,j)}$ and $\mathbf{H}_{(j,:)}$. With $\mathbf{\Sigma}$ present, this can become tricky, so we present a variation on the derivation presented in Erichson et al. (2018) by considering just a row (i) and column (p) of the weighted residual.”

The last comma between W(:,j) and H(i,:) should be eliminated.

If this comment is in reference to Eq. 12 in Section 2.1 (HALS Algorithm), the comma has been removed.

The sizes of the matrices Sigmai, Sigmap and Sigma should be explicitly specified.

To define the sizes of the matrices, the following changes will be made:

Line 30 will be changed to:

“. . . suppose that accompanying the dataset $\mathbf{A}$ is an equally sized $(m \times n)$ matrix $\mathbf{\Sigma}$ . . . ”

Line 233 will be changed to:

“In Eq. (13) and Eq. (14), $\mathbf{\Sigma}_i$ are $\mathbf{\Sigma}_p$ are diagonal matrices (of size m and n respectively) with the diagonal elements corresponding to the elements of the i^{th} row (for $\mathbf{\Sigma}_i$) and p^{th} column (for $\mathbf{\Sigma}_p$) of $\mathbf{\Sigma}$.”

The use of italics for Tr and arg is not consistent with the use of romans for min and max. In mathematical typesetting it is better to keep these non-italicized/sans serif.

Notation for trace has been corrected from Tr to Tr in lines 233, 609, 617 and Equations 15, 16, 17, 18, A4, A5. $argmax_{\mathbf{WH}}$ in Equations 8 and 10 are now presented as $\operatorname{argmax}_{\mathbf{WH}}$. The use of Q_{aux} to denote auxiliary terms in section 2.2 (Rotational Considerations) has been modified to Q_{aux}. In section 4.2 (Simple Case), the use of “test” to in matrix subscripts is also converted to non-italicized text (i.e. $\mathbf{A}_{\text{test}}$)

Further similar changes beyond those listed above have been made to ensure the following of appropriate typesetting conventions in the manuscript.

Notation becomes quite confusing in (27) and beyond.

Equations 32 and 33 now precede Equations 30 and 31 in order to maintain consistency with the ordering of equations 28 and 29 (from which equations

30 to 33 were derived). Further changes to maintain notation consistency and increase clarity in equations throughout the manuscript are addressed in other comments.

On p11, last line, it should be "Thus, the algorithm cannot handle ..." to make it clear that you are referring to the algorithm anot not to the uncertainty matrix.

The comment refers to the second sentence of Section 2.3 (External Weighting). The wording has been modified as suggested.

On p12, "denotes the pseudoinverses" should be "denote the pseudoinverses".

The line in question in also in Section 2.3 (External Weighting) and has been modified as suggested.

It is stated that "Mathematically, these methods ..." - state which are these methods.

This comment is in reference to a sentence (line 305) in Section 2.3 (External Weighting). The sentence in question is referring to two sets of equations presented in Equations 35 and 36. The sentence will be modified to the following for clarity:

"Mathematically, the two methods for calculating **W** and **H** detailed in Equations 35 and 36 respectively are identical, as long as the rank of the factor matrices is equal to k ..."

On p12, line 312 and beyond, the initialization of H is not clear (also better say "initialize" rather than "initiate"). In particular, formula (37) shows H_0 on the left side and H on the right. How is H chosen on the right side?

The wording has been modified from "initiate" to "initialize" as suggested. Equation 37 and the text above it describing the appropriate scaling factor have been modified to the following to remove use of $mean()$ to denote the element-mean of a matrix and also to include $\hat{\mathbf{H}}$ on the right hand side of the equation (as opposed to $\mathbf{H}$ in previous versions of the manuscript):

$$\mathbf{H}_0 = \left(\sqrt{\frac{\bar{\mathbf{A}}}{k\bar{\mathbf{H}}^2}} \right) \hat{\mathbf{H}} \tag{3}$$

In Eq. (3), $\bar{\mathbf{A}}$ denotes the element-mean of matrix $\mathbf{A}$ and $\bar{\mathbf{H}}$ denotes the element-mean of $\hat{\mathbf{H}}$. As $\hat{\mathbf{H}}$ is defined earlier in line 297, this should be sufficient to clarify the initialization of $\mathbf{H}_0$.

On p12, line 319 and possibly elsewhere, the font for H and W is inconsistent (non-bold) with the (bold) font for the same variables elsewhere in the text.

Notation has been standardized to utilize bold font for all matrices and bold, italic font for vectors.

When defining norms it is better to use some space, a variable symbol or a dot instead of using the two double vertical bars next to each other.

Matrix norms (previously denoted $||||_F$ in the case of the Frobenius norm, for example) have been changed to $||\cdot||_F$

In line 377, it is stated that MU is initialized with random numbers. I assume this means that the W and H factors in MU are initialized with nonnegative random values. If this is so, it needs to be said. The distribution also needs to be mentioned (e.g. are they uniformly distributed pseudorandom numbers in [0,1]?)

This comment refers to the second paragraph of the Results section. The paragraph will be modified as follows:

"In the MU algorithm, the elements of $\mathbf{W}$ and $\mathbf{H}$ are initialized by taking the absolute value of random numbers drawn from a standard normal distribution. These values are then scaled by a factor of $\bar{\mathbf{A}}^2$. The ALS and HALS algorithms are initialized by the nonnegative double SVD (NNDSVD) approach detailed in Boutsidis and Gallopoulos (2008)."

The quantity $\bar{\mathbf{A}}$ is previously-defined earlier in the manuscript to represent the element-mean of the matrix $\mathbf{A}$.

The authors still make the error and use the term "non-negative SVD" in line 379, in contrast with the term used for NNDSVD in line 378. See also the comment labeled [page13] in my original review.

The instance of "nonnegative SVD" mentioned by the reviewer has been corrected to "NNDSVD" (previously defined as nonnegative double SVD).

It is also hard not to notice that the authors write "nonnegative" in lines 377-8 and "non-negative" in line 379.

Instances of words prefaced by the prefix "non" (e.g. "non-negative") have been changed such that they are unhyphenated (e.g. "nonnegative"). It is possible though that LaTeX is hyphenating the words if they wrap around to a new line.

On p. 8, the W(:,j), H(j,:) are introduced as the j-th column and row of W and H, while in line 394, the notation becomes W_j and H_j. Incidentally, if you follow the parenthesized notation, then element at row-i, column-j should be denoted by W(i,j) and not W_{ij}.

All equations have been modified to utilize the following notational conventions:

- $\mathbf{W}_{(:,j)}$ and $\mathbf{H}_{(j,:)}$ denote the j^{th} column and row of matrices $\mathbf{W}$ and $\mathbf{H}$ respectively
- When necessary, $\mathbf{R}_{j(i,:)}$ is used to denote the i^{th} row of the j^{th} residual matrix ($\mathbf{R}$)

- $\mathbf{W}_{ij}$ denotes the element in row i and column j in matrix $\mathbf{W}$. This is chosen in contrast with the reviewer's suggestion that $\mathbf{W}_{(i,j)}$ be used to denote the element of matrix $\mathbf{W}$ in order to reduce clutter within equations (and $\mathbf{W}_{ij}$ is common, accepted notation)

2.7 Paragraph 7 and 8

Concerning the bibliography: Some of the references are not suitable. For example: Why refer to Tan et al. 2018 and Takacs and Tikk for HALS? Have these papers introduced HALS? To my understanding, they are only using some version of HALS for some application. Why not simply refer to the primary sources for HALS or even better references that provide a solid foundational discussion, e.g. the monograph by Gillis or the book by Cichocki, Zdunek et al. I would also assume that there must be some specific journal bibliography style.

These references have been removed in favor of the book by Cichocki et al. (2009) as suggested by the reviewer. Formatting in the bibliography has also been addressed.

Several references are incomplete or shown in an inconsistent manner. For example: The Ph.D. thesis by Yahaya does not show the institution. Some wordings are capitalized for no good reason (the institution for Ho's thesis). Using url citations (e.g. Burred J.) that have not undergone proper peer review or at least are posted on a reputable archive (e.g. ArXiv) is not good practice. For this particular one, either the Gillis monograph or the aforementioned monograph by Cichocki, Zdunek et al. noted earlier would be sufficient and more appropriate.

Formatting and content of the references have been updated for completeness and presentation. The reference by Burred has been removed and replaced with peer-reviewed sources as suggested by the reviewer.

References

Boutsidis, C. and Gallopoulos, E.: SVD based initialization: A head start for nonnegative matrix factorization, Pattern Recognition, 41, 1350–1362, https://doi.org/10.1016/j.patcog.2007.09.010, 2008.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S.: Alternating Least Squares and Related Algorithms for NMF and SCA Problems, chap. 4, pp. 203–266, John Wiley Sons, Ltd, ISBN 9780470747278, https://doi.org/10.1002/9780470747278.ch4, 2009.

Erichson, N. B., Mendible, A., Wihlborn, S., and Kutz, J. N.: Randomized Nonnegative Matrix Factorization, Pattern Recognition Letters, 104, 1–7, https://doi.org/10.1016/j.patrec.2018.01.007, 2018.

Massoli, P., Stark, H., Canagaratna, M. R., Krechmer, J. E., Xu, L., Ng, N. L., Mauldin, R. L., Yan, C., Kimmel, J., Misztal, P. K., Jimenez, J. L.,

Jayne, J. T., and Worsnop, D. R.: Ambient Measurements of Highly Oxidized Gas-Phase Molecules during the Southern Oxidant and Aerosol Study (SOAS) 2013, ACS Earth and Space Chemistry, 2, 653–672, https://doi.org/10.1021/acsearthspacechem.8b00028, 2018.

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmospheric Chemistry and Physics, 9, 2891–2918, https://doi.org/10.5194/acp-9-2891-2009, 2009.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Ulbrich, I. M., Ng, N. L., Worsnop, D. R., and Sun, Y.: Understanding Atmospheric Organic Aerosols via Factor Analysis of Aerosol Mass Spectrometry: A Review, Analytical and Bioanalytical Chemistry, 401, 3045 – 3067, 2011.