Response to Review 1

Benjamin Sapper[1], Daven Henze[1], and Jose Jimenez[2]

[1]University of Colorado Boulder, 11 Engineering Dr, Boulder, CO 80309, United States
[2]Cristol Chemistry and Biochemistry, Boulder, CO 80309, United States

Correspondence: Benjamin Sapper (bsapper77@gmail.com)

We thank Anonymous Reviewer 1 for their response. We have considered both reviewers' comments and believe that we have greatly improved the manuscript from them. We also have made additional changes to the manuscript separate from the reviewers' suggestions, mainly making our figures (1, 3, 5-9, 12, 14, 15, A2-A10) clearer, clarifying parts of the text, and correcting spelling and grammar errors. Notably, the external vs. internal weighting comparison in Figures 7, 8, A2, A3, A4, and A5 were redone with a different random seed. However, the interpretation of these results did not change.

In our response, red text is the original comment from the reviewer, black text is our response, and blue text is our updates to the manuscript.

1 Point 1

Could the randomized strategy be explained as a stochastic minimization approach for imposing rank constraints on the NMF solution? Specifically, can it be demonstrated that the NMF solution combined with random projection minimizes a certain cost function? This information would help in evaluating the convergence properties of the proposed method.

The external weighting approach laid out in this paper does not minimize a single cost function, rather two in secession: first the expression

$$||A \oslash \Sigma - \tilde{W}\tilde{H}||_F^2 + \mathcal{L}(\tilde{W}, \tilde{H}) \tag{1}$$

where $\oslash$ is elementwise division, and $\mathcal{L}(\tilde{W}, \tilde{H})$ is a function of regularization terms, and then the expression

$$||(\tilde{W}\tilde{H}) \odot \Sigma - WH||_F^2 + \mathcal{L}(W, H) \tag{2}$$

where $\odot$ is elementwise multiplication. The first minimization finds linear factors of the standardized data, and the second minimization attempts to reweight those factors by the uncertainties of the data. The first minimization is a stochastic block coordinate descent (HALS) as laid out in Erichson et al. (2018), and the second minimization follows alternating least squares (ALS). HALS will converge to a stationary point if negative values are set to a value $\epsilon > 0$, while the latter nonnegative ALS method is not guaranteed to converge (Gillis, 2020). We choose not to add any comments to the manuscript about the theoretical convergence of these algorithms, as that is not the objective of this paper. We update the paper as follows on lines 527-529:

We see that different initializations can lead to different solutions, in terms of both similarity to the given PMF2 solution and weighted error, suggesting that convergence to a global minima isn't always achieved. This further emphasizes the importance of using multiple initializations in order to find an optimal solution.

2 Point 2

Regarding the potential drawback of imposing rank constraints on the NMF outputs, is it feasible to relax these constraints, perhaps by employing the nuclear norm?

Nuclear norm regularization is an increasingly popular tool in matrix factorization as it is the convex envelope to minimizing the rank function of a matrix approximation (Hu et al., 2013). Cai et al. (2010) showed that the unconstrained minimization problem

$$argmin_{\mathbf{X}} \frac{1}{2}||\mathbf{A}-\mathbf{X}||_F^2 + \tau||\mathbf{X}||_*^2 \tag{3}$$

where

$$||\mathbf{X}||_* = \sum_{i=1}^{n} \sigma_i(\mathbf{A})$$

with $\sigma_i(\mathbf{A})$ denoting the i^{th} singular value of $\mathbf{A}$, is solved as $\mathbf{X} = \mathcal{D}_\tau(\mathbf{A})$, which is defined as

$$\mathcal{D}_\tau(\mathbf{A}) = \mathbf{U}(\mathbf{\Sigma} - \tau\mathbf{I})_+\mathbf{V}^T \tag{4}$$

where $\mathbf{U\Sigma V}^T$ is the singular value decomposition of $\mathbf{A}$, and $(\mathbf{M})_+$ projects all negative values in $\mathbf{M}$ to zero. Naturally, as τ increases, the solution $\mathbf{X} = \mathcal{D}_\tau(\mathbf{A})$ decreases in rank.

Two possible ways to apply nuclear norm regularization to weighted PMF are by the Alternating Direction Method of Multipliers (ADMM), as demonstrated in Sun and Mazumder (2013), and by reconstruction of the nuclear norm into a Frobenius norm, which can then be treated as L2 regularization (Fornasier et al., 2011). However, both these approaches involve key computations with the large low rank estimate $\mathbf{X}$, and thus may be computationally expensive to implement. Furthermore, both algorithms still involve some arbitrary choice of rank by requiring a preset amount of nuclear norm regularization τ. In traditional PMF, factor profiles are optimized from a prechosen amount of factors, and thus the we centered our paper around this approach. We update the manuscript on lines 356-367 as:

An increasingly popular alternative to traditional regularization is nuclear norm regularization, which can be applied to matrix factorization without the need for rank constraints (Hu et al., 2013; Sun and Mazumder, 2013; Fornasier et al., 2011). The nuclear norm is defined as

$$||\mathbf{X}||_* = \sum_{i=1}^{n} \sigma_i(\mathbf{A})$$

where $\sigma_i(\mathbf{A})$ is the i^{th} singular value of $\mathbf{A}$. It is the convex envelope to the rank function, and thus finding a PMF solution that minimizes the nuclear norm also has a minimal rank (Hu et al., 2013). Two possible ways to apply nuclear norm regularization to weighted PMF are by the Alternating Direction Method of Multipliers (ADMM), as demonstrated in Sun and Mazumder (2013), and by reconstruction of the nuclear norm into a Frobenius norm, which can then be treated as L2 regularization (Fornasier et al., 2011). However, both these approaches involve key computations with the large low rank product **WH**, and thus may be computationally expensive to implement. Furthermore, both algorithms still involve some arbitrary choice of rank by requiring a preset amount of nuclear norm regularization. In traditional PMF, factor profiles are optimized from a prechosen amount of factors, and thus the we center our results around this approach.

3 Point 3

Lastly, the manuscript does not do a good job of citing the related state-of-the-art methods. For example, there has been a tremendous amount of work done in relation to randomized weighted NMF. Please see the following:

Yahaya, F., Puigt, M., Delmaire, G., Roussel, G. (2021, June). Random Projection Streams for (Weighted) Nonnegative Matrix Factorization. In IEEE ICASSP 2021.

Yahaya, F. (2021, November). Compressive informed (semi-) non-negative matrix factorization methods for incomplete and large-scale data: with application to mobile crowd-sensing data. Université du Littoral Côte d'Opale.

Yahaya, F., Puigt, M., Delmaire, G., Roussel, G. (2020). Gaussian Compression Stream: Principle and Preliminary Results. arXiv preprint arXiv:2011.0539.

The authors recognize that work done by Dr. Yahaya and their colleagues was understated in the paper. Thus, we have added two subsections to our manuscript, one in the External Weighting subsection of the Background Section (lines 163-180), and one in the Results Section under subsection 4.3 (lines 482-519). We have also added Table 1 (Table 3 in the revised manuscript).

3.1 1.5.1 Expectation Maximization

The Expectation Maximization (EM) approach was first designed for matrix factorization problems associated with missing entries (Zhang et al., 2006). Specifically, if A^o is the observed data and A^u is the unknown data within **A**, then the EM approach seeks to find factors **W** and **H** that satisfy the following (Zhang et al., 2006)

$$argmax_{\mathbf{WH}}\mathbb{E}(\log(\mathbb{P}(A^o, A^u|\mathbf{WH}))|A^o, \mathbf{WH}^{(t-1)}) \tag{5}$$

where $\mathbf{WH}^{(t-1)}$ is the product of the previous estimates of $\mathbf{W}$ and $\mathbf{H}$, and $\mathbb{P}$ is a probability measure. This problem is equivalent to running a PMF algorithm on the following adjusted matrix (Zhang et al., 2006)

$$\mathbf{A}_1 = \mathbf{C} \odot \mathbf{A} + (\mathbf{1} - \mathbf{C}) \odot \mathbf{WH}^{(t-1)} \tag{6}$$

with $\mathbf{C}_{ij} = 1$ if $\mathbf{A}_{ij}$ is known and $\mathbf{C}_{ij} = 0$ if $\mathbf{A}_{ij}$ is unknown, and $\mathbf{1}$ is a matrix of ones.

Recent work has looked into expanding on this approach to continuous weights as seen in most PMF problems of aerosol data (Yahaya et al., 2019; Yahaya, 2021; Yahaya et al., 2021). To handle the continuous case, a variation of Eq. 5 is maximized (Yahaya et al., 2019)

$$argmax_{\mathbf{WH}} \mathbb{E}(\log(\mathbb{P}(\mathbf{C} \odot \mathbf{A}, (\mathbf{1} - \mathbf{C}) \odot \mathbf{A}_{theo} | \mathbf{WH})) | \mathbf{C} \odot \mathbf{A}, \mathbf{WH}^{(t-1)}) \tag{7}$$

where $\mathbf{C}$ is now a weight matrix containing estimates of confidence as a value between 0 and 1 in a given data point, and $\mathbf{A}_{theo}$ is the theoretical true data. Maximizing Eq. 7 is equivalent to running any PMF algorithm on the matrix $\mathbf{A}_1$ formed in Eq. 6.

How should one form the confidence matrix $\mathbf{C}$ from the uncertainties matrix $\mathbf{\Sigma}$? Yahaya (2021) suggests in scaling the weights (in this case $1/\sigma_{ij}$) so that the maximum value is 1. However, previous testing has primarily focused on problems with binary weights (Yahaya et al., 2019; Yahaya, 2021; Yahaya et al., 2021).

3.2 4.3.2 Comparison Between Expectation Maximization and External Weighting

To test the Expectation Maximization (EM) approach to uncertainties weighting as mentioned in Section 1.5.1, the weights σ_{ij} in the uncertainties matrix Σ are scaled so that $\max_{i,j} 1/\sigma_{ij} = 1$. Since the bulk computational component of the algorithm is constructing the matrix $\mathbf{A}_1$ in Eq. 6, the authors in Yahaya et al. (2019) and Yahaya et al. (2021) recommend updating $\mathbf{A}_1$ only after convergence or a maximum number of iterations of 20 or 50. They also note that applying the expectation step too early in the algorithm led to poorer performance due to the amount of error in the estimates of $\mathbf{W}$ and $\mathbf{H}$. Since we wanted to apply the same stopping criterion as mentioned earlier, we chose to reconstruct $\mathbf{A}_1$ at fixed iterations - after 10 and 20 PMF steps for different experiments. The first construction of $\mathbf{A}_1$ is done at the 1^{st}, 5^{th}, and 10^{th} step. We used the NNDSVD initialization in (Boutsidis and Gallopoulos, 2008), and also varied the tolerance of the stopping condition between 10^{-4} and 10^{-6}. We summarize these results by presenting the range of average values across the different variations.

We present a comparison of external weighting and the Expectation Maximization algorithm in Table 1, using the ranges of values from the different experiments listed above. Each value is an average over 20 trials. We compare the convergence time of the algorithm, the number of steps, the weighted residual error, and the similarity to the PMF2 solution for both $\mathbf{W}$ (correlation coefficient) and $\mathbf{H}$ (cosine similarity).

As demonstrated in Table 1, some variations of the EM algorithm were able to outperform externally weighted HALS and RHALS in total time, as well as in weighted error. For instance, running the expectation maximization algorithm with the first calculation of $\mathbf{A}_1$ taking place at the 10^{th} step, and recalculating $\mathbf{A}_1$ after 20 additional steps, until the convergence criterion with a tolerance of 10^{-5} was reached, yielded an average weighted error of 6.92×10^3 in 0.6703 seconds. This computational speed compares to RHALS, while the accuracy bests both RHALS and externally weighted HALS. However, no expectation

Table 1. Average statistics of external weighting (EW) versus Expectation Maximization (EM) algorithms over 20 trials. Internally weighted (IW) HALS is provided as a reference. Ranges of the values of the algorithms run using different variations expectation maximization (EM) steps are presented. The correlation of the columns of **W** and similarity of the rows of **H** to the PMF2 solution are also listed.

Comparison of Algorithms					
Algorithm	Total Time (s)	Steps	Weighted Error	Correlation of **W**	Similarity of **H**
HALS (IW)	40.70	19.65	6.46×10^3	0.8756	0.9134
HALS (EW)	1.07	31.80	7.09×10^3	0.8414	0.8828
RHALS (EW)	0.56	38.35	7.27×10^3	0.8475	0.9034
HALS (EM)	0.51-2.93	13.50-83.90	$6.71\text{-}7.22\times10^3$	0.7759-0.8306	0.8434-0.8861
RHALS (EM)	0.33-1.19	20.30-61.05	$7.11\text{-}7.54\times10^3$	0.7930-0.8221	0.8301-0.8694

maximization algorithm was as successful at recreating the PMF2 time series factors, as seen in column 5 of Table 1. Externally weighted RHALS and HALS also provided mass spectra factors with a higher similarity to the PMF2 factors, with the exception of two runs of the expectation maximization algorithm, one recalculating $\mathbf{A}_1$ after 20 steps, with the first calculation at the 5^{th} step, and the other after 10 steps, with the first calculation at the 1^{st} step. These yielded average similarities of 0.8856 and 0.8861, respectively, although they both took over 2.90 seconds to run, much slower than any other algorithm tested besides internally weighted HALS.

We also tested how well each algorithm produced factors that were within a rotation of the PMF2 factors. As detailed in Section 1.4, the factor profiles of $\hat{\mathbf{W}}\hat{\mathbf{H}} = (\mathbf{W}\mathbf{T}^{-1})(\mathbf{T}\mathbf{H})$ for a square matrix $\mathbf{T}$ may be closer to the desired solution than the original factors $\mathbf{W}$ and $\mathbf{H}$. To see the extent that $\mathbf{W}$ can be rotated towards $\mathbf{W}_{\text{PMF2}}$, we find the matrix $\mathbf{T}$ that minimizes the total squared difference between $\mathbf{H}_{\text{PMF2}}$ and $\mathbf{T}\mathbf{H}$, and then find the average correlation between the rows of $\mathbf{W}_{\text{PMF2}}$ and $\mathbf{W}\mathbf{T}^{-1}$. A symmetrical approach can be made to find the cosine similarity for $\mathbf{H}$. For internally weighted HALS, externally weighted HALS, and RHALS, we found average post rotation time series correlations to be 0.9391, 0.9021, and 0.8902, respectively, and post rotation mass spectra similarities to be 0.9602, 0.9433, and 0.9518, respectively. When this approach was tested on HALS and RHALS using the EM algorithm, average post rotation time series correlations varied within 0.8528-0.8819 and 0.8314-0.8544, respectively, and post rotation mass spectra similarities within 0.9230-0.9363 and 0.9147-0.9287, respectively.

Ultimately, we found that external weighting recreated the PMF2 factors more consistently than expectation maximization, both before and after rotation. This may be due to the fact that the scaling of the weights used for the EM step is not perfectly analogous to creating a set of weights that represent the confidence of each data point. Thus, the EM method using this scaling may not capture the key error weighted patterns in the data as well as external weighting.

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.

Cai, J.-F., Candès, E. J., and Shen, Z.: A Singular Value Thresholding Algorithm for Matrix Completion, SIAM Journal on Optimization, 20, 1956–1982, https://doi.org/10.1137/080738970, 2010.

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.

Fornasier, M., Rauhut, H., and Ward, R.: Low-rank Matrix Recovery via Iteratively Reweighted Least Squares Minimization, SIAM Journal on Optimization, 21, 1614–1640, https://doi.org/10.1137/100811404, publisher: Society for Industrial and Applied Mathematics, 2011.

Gillis, N.: Chapter 8: Iterative algorithms for NMF, in: Nonnegative Matrix Factorization, Data Science, pp. 261–305, Society for Industrial and Applied Mathematics, https://doi.org/10.1137/1.9781611976410.ch8, 2020.

Hu, Y., Zhang, D., Ye, J., Li, X., and He, X.: Fast and Accurate Matrix Completion via Truncated Nuclear Norm Regularization, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, 2117–2130, https://doi.org/10.1109/TPAMI.2012.271, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

Sun, D. L. and Mazumder, R.: Non-negative matrix completion for bandwidth extension: A convex optimization approach, in: 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6, https://doi.org/10.1109/MLSP.2013.6661924, iSSN: 2378-928X, 2013.

Yahaya, F.: Compressive informed (semi-)non-negative matrix factorization methods for incomplete and large-scale data : with application to mobile crowd-sensing data, Ph.D. thesis, 2021.

Yahaya, F., Puigt, M., Delmaire, G., and Roussel, G.: How to Apply Random Projections to Nonnegative Matrix Factorization with Missing Entries?, in: 2019 27th European Signal Processing Conference (EUSIPCO), pp. 1–5, https://doi.org/10.23919/EUSIPCO.2019.8903036, iSSN: 2076-1465, 2019.

Yahaya, F., Puigt, M., Delmaire, G., and Roussel, G.: Random Projection Streams for (Weighted) Nonnegative Matrix Factorization, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3280–3284, https://doi.org/10.1109/ICASSP39728.2021.9413496, iSSN: 2379-190X, 2021.

Zhang, S., Wang, W., Ford, J., and Makedon, F.: Learning from Incomplete Ratings Using Non-negative Matrix Factorization, in: Proceedings of the 2006 SIAM International Conference on Data Mining (SDM), Proceedings, pp. 549–553, Society for Industrial and Applied Mathematics, https://doi.org/10.1137/1.9781611972764.58, 2006.