Nonnegative Matrix Factorization with Group and Basis Restrictions

Phillip Shreeves¹,³ · Jeffrey L. Andrews¹ · Xinchen Deng² · Ramie Ali-Adeeb² · Andrew Jirasek²

Received: 10 November 2022 / Revised: 6 October 2023 / Accepted: 8 October 2023 / Published online: 4 November 2023
© The Author(s) under exclusive licence to International Chinese Statistical Association 2023

Abstract
Nonnegative matrix factorization (NMF) is a popular method used to reduce dimensionality in data sets whose elements are nonnegative. It does so by decomposing the data set of interest, X, into two lower rank nonnegative matrices multiplied together. These two matrices can be described as the latent factors, represented in the rows of H, and the scores of the observations on these factors that are found in the rows of W. This paper provides an extension of this method which allows one to specify prior knowledge of the data, including both group information and possible underlying factors. This is done by further decomposing the matrix, H, into matrices A and S multiplied together. These matrices represent an ‘auxiliary’ matrix and a semi-constrained factor matrix, respectively. This method and its updating criterion are proposed, followed by its application on both simulated and real-world examples displaying different uses of the algorithm.

Keywords Nonnegative matrix factorization · Semi-supervised learning · Raman spectroscopy · Dimensionality reduction

Phil_Shrreeves
phil_shreeves@sfu.ca
Jeffrey L. Andrews
jeff.andrews@ubc.ca
Andrew Jirasek
andrew.jirasek@ubc.ca

¹ Department of Statistics, University of British Columbia - Okanagan, Kelowna, British Columbia, Canada
² Department of Physics, University of British Columbia - Okanagan, Kelowna, British Columbia, Canada
³ Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada

Springer
1 Introduction

Dimensionality reduction is an important tool in fighting the curse of dimensionality. This is typically done by mapping the data to a lower dimensional subspace in some way, while also retaining as much information as possible from the original data set. There is a large number of techniques that are able to perform dimensionality reduction in this manner, including principal component analysis [PCA, [1]], vector quantization [VQ, [2]], factor analysis [3], and nonnegative matrix factorization [NMF, [4]]. All of these techniques follow two key properties of dimensionality reduction specified by Wang and Zhang [5], which includes (1) reducing the total number of dimensions in the data and (2) recovering factors in the data that can be interpreted effectively.

NMF differs from the other methods specified above primarily due to the fact that it has the added constraint of nonnegativity throughout its matrices. It also benefits as the recovered components represent partitioned regions of the observations, as opposed to reduced amounts of an entire observation, which can be beneficial in many applications. For example, Lee and Seung [4] first introduce NMF by applying it to a data set containing facial images. NMF is able to reduce the dimensionality by identifying components such as the nose, eyes, and mouth. PCA, on the other hand, identified entire facial structures as components, making it difficult to envision the observations as a linear combination of parts. We propose a restricted algorithm that allows for one to specify expected factors and groupings as added constraints to the model, leading to a more flexible modeling paradigm for the user.

The proposed algorithm is particularly useful in the field of Raman spectroscopy (RS), an optical interrogation method whereby vibrational modes of constituent molecules are identified through inelastic light scattering. Raman spectra can provide detailed information on a range of molecular constituents within a single sample acquisition [6–8]. Relative changes in peak areas and/or positions can be used to identify altered molecular dynamics within a system undergoing a given perturbation. Typical dimensionality reduction practice in the field of RS is to use principal component analysis [9–12]. However, this is not an ideal model as the constituent underlying spectra should not be permitted to take on negative values. Furthermore, it is inappropriate to make the assumption that all constituents, which in this case are represented through the principal components, are orthogonal. This is because multiple Raman bands typically originate from a single chemical species. Therefore, one could expect these bands to be correlated in their response to a given perturbation. Thus other dimensionality reduction techniques, such as NMF, are more appropriate for application to these data sets. In the following sections, the NMF model is extended to improve overall interpretation and allow for suspected chemical constituent spectra to be specified during model fitting. Other models are also applied to these data in order to further expand our knowledge in the field of RS.
2 Background

2.1 Nonnegative Matrix Factorization

Nonnegative matrix factorization was originally developed by Lee and Seung [4] as a method of finding the major underlying factors of a data set and determining how each individual observation is approximated using a linear combination of these factors. The factors are found by mapping the observations to a latent factor space that accounts for the vast majority of variance in the model. Each observation is then written as a linear combination of the discovered factors. These factors and coefficients of the linear combinations are found by decomposing the nonnegative data of interest $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times p}$ (where $n$ is the number of observations and $p$ is the number of variables) into two lower rank nonnegative matrices $\mathbf{W} \in \mathbb{R}_{\geq 0}^{n \times q}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{q \times p}$ such that

$$\mathbf{X} \approx \mathbf{WH},$$

where $q$ is the number of underlying factors. We can describe the span of the rows of $\mathbf{H}$ as the $q$-dimensional subspace of feasible observations. As such, the rows of $\mathbf{W}$ contain coefficients describing how each observation is mapped to the subspace. Thus, each observation $\mathbf{x}_i$ can be viewed as an additive linear combination of the features contained in the rows of $\mathbf{H}$ that are weighted by the scores contained in the columns of $\mathbf{W}$. This equation is an approximation of the assumed true decomposition of $\mathbf{X}$,

$$\mathbf{X} = \mathbf{WH} + \mathbf{\varepsilon},$$

where $\mathbf{\varepsilon} \in \mathbb{R}^{n \times p}$ represents the residual or noise matrix of the model. This noise $\mathbf{\varepsilon} = \mathbf{X} - \mathbf{WH}$, is a key component in the model as its assumed probability distribution assists in determining what objective function the data need to be optimized with respect to [5]. For example, assuming Gaussian distributed noise, the Frobenius norm

$$D_F(\mathbf{X} \mid \mathbf{WH}) = \frac{1}{2} \| \mathbf{X} - \mathbf{WH} \|_F^2 = \frac{1}{2} \sum_{ij} (x_{ij} - [WH]_{ij})^2$$

is the function that is to be optimized [5, 13]. Another function commonly used is the generalized Kullback–Leibler divergence (GKLD) which is used when assuming the data arise from a Poisson process [5, 13, 14]. In this paper, Gaussian noise will be assumed as the simulated data sets were created with this noise and preliminary diagnostics regarding real data sets for complex Raman spectra (such as those in Sect. 4.3) suggest this may be a reasonable assumption.

Lee and Seung [13] describe that NMF can be formulated by minimizing the function $\| \mathbf{X} - \mathbf{WH} \|^2$. While this function is convex in $\mathbf{W}$ or $\mathbf{H}$ individually, they are unfortunately not convex in both variables together. Finding global minima is not guaranteed [15] because of this, and algorithms which find local minima are predominantly considered at the moment. Additionally, solutions found using NMF are typically non-unique, a property studied in depth by Donoho [16]. The prototypical
algorithm for finding these matrices is an alternating multiplicative updating process that uses the following updates:

\[ h_{ij} \leftarrow h_{ij} \frac{(W^T X)_{ij}}{(W^T WH)_{ij}}, \quad w_{ij} \leftarrow w_{ij} \frac{(XH^T)_{ij}}{(WHH^T)_{ij}}. \]  

(4)

This updating process is considered slow by most accounts and scales dramatically in elapsed time when requiring both the rank and dimension of \( X \) to increase [5, 13]. However, this is not a concern with respect to the motivating application found in Sect. 4.3. Suggestions to increase the performance of the algorithm include applying gradient descent algorithms [17], the conjugate gradient [18], or the use of Nestorov’s optimal gradient method [19]. While NMF is advantageous due to its added interpretability in comparison with PCA, it does fall victim to slower computation speed [20]. Aside from the assumption of nonnegativity, Lee and Seung [4] also note that no further assumptions are made about the statistical independencies of the underlying factors.

### 2.1.1 Extensions of Nonnegative Matrix Factorization

Various authors have considered constraints to address questions concerning clustering and classification with the NMF algorithm. Ding et al. [21] suggest that when the Frobenius norm is used as the objective function, the standard NMF algorithm bears resemblance to a relaxed form of K-means clustering [22] where one factor matrix contains the centroids of the clusters and the other contains indicators of cluster memberships. This is then expanded upon by Li et al. [23] in a semi-supervised clustering manner, where limited knowledge of cluster membership is present. Wagstaff et al. [24] specify how semi-supervised clustering is an algorithm performed with respect to two separate constraints. These two constraints are that (1) there is the existence of must-link observations where the observations must be grouped into the same cluster and (2) there also exists cannot-link observations that are required to be in separate clusters.

When this knowledge is limited to a number of observations \( l < n \), a semi-supervised clustering algorithm is needed. Liu et al. [25] address these constraints by introducing the decomposition of the feature scores matrix \( W \) into two lower rank matrices, consisting of a constrained grouping matrix as well as an ‘auxiliary matrix.’ Here, the grouping matrix uses the \( a \) priori label information and stores it in the form of an indicator matrix \( C = [c_1 \ c_2 \ \cdots \ c_g] \) where \( g \) is the total number of classes. This matrix is then expanded upon with each unsupervised observation being assigned its own class such that \( W = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \). Both \( W \) and the auxiliary matrix are then updated iteratively with that information taken into account. While this algorithm serves a different purpose than the one that will be described in Chapter 3, it still provides a contribution to the updating process described in Sect. 3.2. Yang et al. [26] further expand upon the work of Liu et al. [25] by adding in a sparsity constraint, making it a dually constrained algorithm.
Arora et al. [27] introduced the concept of separable nonnegative matrices, which are computed more efficiently, even in the presence of noise. A nonnegative matrix, $X$, is defined to be separable if there exists an index set $K$ of cardinality $r$ and a nonnegative matrix $H$ such that $X = X(:, K)H$. This concept was then expanded upon by Gillis [28] to create the Successive Nonnegative Projection Algorithm, a fast algorithm similar to NMF that uses near-separable matrices in order to decompose the data of interest. This algorithm, as well as separable nonnegative matrix factorization [29], has been used in multiple spectral data applications [30, 31].

Nonnegative matrix tri-factorization, a method which involves the decomposition of the data matrix into three lower rank matrices, has also become an increasingly popular research topic. While unconstrained 2-factor NMF and unconstrained 3-factor NMF have been found to be equivalent, constrained versions tend to differ [32]. This is because the added matrix provides additional degrees of freedom, which accommodates the constraints imposed on the other matrices. Ding et al. [32] introduce orthogonal nonnegative matrix tri-factorization, which imposes orthogonality constraints on the two outside matrices. These constraints allow for one to simultaneously cluster both the rows and columns of the original matrix of interest, $X$. Chen et al. [33] utilize this method to aid in collaborative filtering aiming to predict a user’s ratings on items according to a collection of other like-minded users’ ratings information. In a similar application, Li et al. [34] extend upon this method by incorporating label information for sentiment analysis.

3 Methodology

3.1 The Group and Basis-Restricted NMF (GBR-NMF) Algorithm

The conventional NMF decomposition proposed by Lee and Seung [4] is computed using a completely unsupervised algorithm, meaning one cannot specify any suspected factors in the model or specify any clustering constraints if there are suspected groups within the data. Here, we propose a variation of the work by Liu et al. [25] by similarly utilizing a further decomposition of the data of interest into matrices $W$, $A$, and $S$ such that

$$X = WAS + \epsilon,$$

where it is apparent that $H$ from Eq. 1 is further decomposed into a $q \times q$ auxiliary matrix $A$ and the $q \times p$ matrix $S$ containing the underlying factors. However, this decomposition differs in the way that restrictions are imposed on the matrices. Liu et al. [25] restrict their $W$ matrix as stated in Sect. 2.1.1, whereas the new algorithm restricts a number of columns in $W$ through the use of a complete indicator matrix and also restricts some rows in $S$. The added restrictions allow for one to specify known groups in $W$ and suspected factors in $S$, further improving the interpretability of the model. It is important to note that while there is further decomposition, the matrices $W$ and $S$ can still be interpreted the same way as $W$ and $H$, respectively. However, due to the fact that we will now be imposing a constraint on $S$, $A$ is
needed to perform adjustments on the constrained factors. Similar to the third matrix in nonnegative matrix tri-factorization, \( A \) introduces additional degrees of freedom into the decomposition. This allows for constraints to be made on \( W \) and \( S \) without severely compromising the error between the true and reconstructed observations. We propose that \( A \) be initialized as an identity matrix which, through the iterative updating process, will act as a scaling matrix to the rows of \( S \) — a detail which is shown to be quite meaningful in Sect. 4.3.

As previously specified, constraints will now be added to \( S \), herein creating a restricted model. Assuming that there are \( q \) underlying factors in the model, suppose there exist \( k \) known and \( q - k \) unknown factors in the model. Then these \( k \) known factors can be specified as factors in \( S \) that are not updated through the iterative process, while the \( q - k \) factors are still included in the updating process. This allows for the algorithm to take into consideration the factors that a user may know to be present and use them to estimate other factors that may not be known to the users.

Branching from the work of Ding et al. [21], an optional constraint regarding the \( W \) matrix has also been added for applications regarding clustering. As previously described in Sect. 2.1.1, basic NMF resembles a soft K-means clustering algorithm under the use of the Frobenius norm as an objective function. Here, the features in \( S \) represent the centroids in the algorithm and the scores in \( W \) are how each observation scores with respect to these centroids. With this in mind, we have allowed the constraint of the first \( g \) columns of \( W \) to be a known grouping matrix. These columns are never updated in the model and the factors regarding these columns are kept unconstrained.

Thus, in the case where both constraints are employed, we have \( g + k \leq q \) with the first \( g \) columns of \( W \) and rows of \( S \) responsible for the clustering results with the next \( k \) accounting for the constrained factors. This allows for the major differences in groups to be found while other dissimilarities are found in the other \( q - g \) factors. We previously noted in Sect. 2 that the conventional NMF algorithm falls victim to non-uniqueness and convergence to local minima with respect to the objective function. While it has been found that these problems can be mitigated by implementing constraints [35], we do not guarantee that they are resolved with GBR-NMF. However, these problems did not arise in any of the applications discussed in Sect. 4.

### 3.2 Updating Process

Assuming we are updating with respect to Gaussian distributed error, we are required to update with respect to the Frobenius norm

\[
D_F(X \mid WAS) = \frac{1}{2} \| X - WAS \|^2
\]

\[\propto \| X - WAS \|^2
\]

(6)
as the objective function of interest with the constraint that \( w_{ij} \geq 0, a_{ij} \geq 0, \) and \( s_{ij} \geq 0 \) for all \( ij \in \mathbb{R} \). As previously stated in Sect. 2.1, it is not feasible to find global minima as GBR-NMF is not convex in \( W, A, \) and \( S \). It instead requires an updating process to find local minima. To ensure a suitable local minimum has been achieved, the GBR-NMF algorithm was applied to different data sets multiple times in order to ensure consistency in the accuracy of the algorithm. These different test cases included different initializations of the \( W, A, \) and \( S \) matrices in an attempt to find different local minima. If one finds themselves in the situation where they believe the local minimum is unsatisfactory, they too could use the algorithm in multiple cases and compare Frobenius norms to find the most suitable answer.

Applying the trace function, as well as both the cyclic and transpose properties associated with it, we have

\[
D(X | WAS) \propto Tr((X - WAS)(X - WAS)^T) \\
= Tr(XX^T) - 2Tr(WASX^T) + Tr(WASS^T A^TW^T).
\]

(7)

However, this does not account for the nonnegativity constraints previously specified. In order to do so, a Lagrange function \( (\mathcal{L}) \) with multipliers \( \alpha_{ij}, \beta_{ij} \) and \( \gamma_{ij} \) is required for constraints \( w_{ij} \geq 0, a_{ij} \geq 0 \) and \( s_{ij} \geq 0 \), respectively (note that \( \alpha = [\alpha_{ij}], \beta = [\beta_{ij}], \) and \( \gamma = [\gamma_{ij}] \)). This gives the following Lagrange function

\[
\mathcal{L} = Tr(XX^T) - 2Tr(WASX^T) + Tr(WASS^T A^TW^T) \\
+ Tr(\alpha W^T) + Tr(\beta A^T) + Tr(\gamma S^T).
\]

(8)

Now needing the derivatives of \( \mathcal{L} \) to be zero with respect to \( W, A, \) and \( S \), we have

\[
\frac{\partial \mathcal{L}}{\partial W} = -2XS^TA^T + 2WASS^TA^T + \alpha = 0,
\]

\[
\frac{\partial \mathcal{L}}{\partial A} = -2WTXS^T + 2WTWASS^T + \beta = 0,
\]

\[
\frac{\partial \mathcal{L}}{\partial S} = -2A^TW^TX + 2A^TW^TWAS + \gamma = 0.
\]

(9)

Then applying the Karush-Kuhn-Tucker conditions \( \alpha_{ij}w_{ij} = 0, \beta_{ij}a_{ij} = 0, \) and \( \gamma_{ij}s_{ij} = 0 \), we obtain

\[
(XS^TA^T)_{ij}w_{ij} - (WASS^TA^T)_{ij}w_{ij} = 0,
\]

\[
(WTXS^T)_{ij}a_{ij} - (WTWASS^T)_{ij}a_{ij} = 0,
\]

\[
(A^TW^TX)_{ij}s_{ij} - (A^TW^TWAS)_{ij}s_{ij} = 0.
\]

(10)

which leads to the following updating rules
Further detail of the derivations in this section can be found in Appendix A.

### 3.3 Proof of Convergence

In order to prove that the above updates are correct, the objective function must be non-increasing according to the updates. The objective function is then invariant under the updates if and only if $W$, $A$, and $S$ are at stationary points [25]. In order to prove this, we must borrow a lemma stemming from updates to the EM algorithm [36, 37] provided below.

**Lemma 3.1** If an auxiliary function, $G$, exists for $F(x)$ and satisfies the conditions

1. $G(x, x') \geq F(x)$
2. $G(x, x) = F(x)$

then $F$ is non-increasing using the update

$$x^{t+1} = \arg \min_x G(x, x')$$

Given that $F_{w_{ij}}$ is defined as the parts of equation 6 relevant to $w_{ij}$ and $F'_{w_{ij}}$ are the derivative of such, we then need to prove the following three lemmas.

**Lemma 3.2** The function

$$G(w, w_{ij}) = F_{w_{ij}} (w'_{ij}) + F'_{w_{ij}} (w_{ij}) (w - w'_{ij}) + \frac{(\text{WASS}^T A^T)_{ij}}{w'_{ij}} (w - w'_{ij})^2$$

is an auxiliary function to $F_{w_{ij}}$. 

$$w_{ij} \leftarrow w_{ij} \frac{(X S^T A^T)_{ij}}{(\text{WASS}^T A^T)_{ij}}, \quad (11)$$

$$a_{ij} \leftarrow a_{ij} \frac{(W^T X S^T)_{ij}}{(W^T \text{WASS}^T)_{ij}}, \quad (12)$$

$$s_{ij} \leftarrow s_{ij} \frac{(A^T W^T X)_{ij}}{(A^T W^T \text{WAS})_{ij}}. \quad (13)$$
Lemma 3.3 The function

\[ G(a, a_{ij}) = F_{a_{ij}}(a_{ij}) + F'_{a_{ij}}(a) + \frac{(W^T W A S S^T)_{ij}}{a_{ij}}(a - a_{ij})^2 \]

is an auxiliary function to \( F_{a_{ij}} \).

Lemma 3.4 The function

\[ G(s, s_{ij}) = F_{s_{ij}}(s_{ij}) + F'_{s_{ij}}(s) + \frac{(A^T W^T W A S S)_{ij}}{s_{ij}}(s - s_{ij})^2 \]

is an auxiliary function to \( F_{s_{ij}} \).

These lemmas are proved in a very similar fashion. As a result, lemma 3.2 will be proved below, while Lemmas 3.3 and 3.4 can be found in Appendix B. With minor modifications of the proof by Liu et al. [25], the following can be proved.

Proof It is clear that \( G(w, w) = F_{w_{ij}}(w) \). According to the definition of an auxiliary function, it only needs to be shown that \( G(w, w_{ij}) \geq F_{w_{ij}}(w) \). This can be done using the Taylor series expansion of \( F_{w_{ij}} \):

\[
F_{w_{ij}}(w) = F_{w_{ij}}(w_{ij}) + F'_{w_{ij}}(w - w_{ij}) + \frac{1}{2} F''_{w_{ij}}(w - w_{ij})^2
\]

with \( F''_{w_{ij}} \) being the second-order derivative of \( F_{w_{ij}} \). We can show that

\[
F_{w_{ij}} = \left( \frac{\partial D}{\partial W} \right)_{ij} = (-2XS^TA^T + 2W A S S A^T)_{ij},
\]

\[
F''_{w_{ij}} = \left( \frac{\partial^2 D}{\partial W^2} \right)_{ij} = 2(\text{ASS}^T A)^T)_{ij}.
\]

As \( F''_{w_{ij}} \) has now been solved for, it is equivalent to prove

\[
\frac{(W A S S^T A^T)_{ij}}{w_{ij}} \geq \frac{1}{2} F''_{ij} = (\text{ASS}^T A)^T)_{ij} \tag{14}
\]

in order to reach Lemma 3.2. Now we have

\[
(W A S S^T A^T)_{ij} = \sum_{\ell=1}^{q} (W)_{i\ell} (\text{ASS}^T A^T)_{\ell j}
\]

\[
\geq w'_{ij} (\text{ASS}^T A)_{ij} \tag{15}
\]

which when rearranged proves Eq. 14. □
3.4 Algorithm Summary

For the observed $n \times p$ data matrix, $X$, with $c$ known groups in $W$, $k$ suspected factors in $S$ and $q$ total factors, the algorithm can be summarized using the following steps:

1. If an $n \times c$ cluster membership matrix is specified, set the first $c$ columns of the matrix, $W$, to be this matrix. Initialize the remaining $n \times (q - c)$ elements sampling from the uniform distribution $U(\min(X), \max(X))$. Otherwise, initialize the entire matrix by sampling from the uniform distribution specified above.
2. Initialize $A$ such that it takes the form of a $q \times q$ identity matrix.
3. Initialize the remaining $(q - k - c)$ columns of $S$ sampling from the uniform distribution on the bounds of the known-factor values.
4. Calculate $D_F^{(0)}$ using Eq. 6.
5. Set $z = 1$ and enter the updating process:
   (a) Update the unspecified elements of $W$ using Eq. 11.
   (b) Update all elements of $A$ using Eq. 12.
   (c) Update the unspecified elements of $S$ using Eq. 13.
   (d) Calculate $D_F^{(i)}$ using equation 6.
   (e) If $D_F^{(z-1)} - D_F^{(z)}$ is less than a specified $\delta$ value or the value of $i$ is equal to the maximum iteration count, move to step 6. Otherwise set $z = z + 1$ and return to step 5a.

6. Return the updated values of $W$, $A$, and $S$.

It is important to note that stopping criterion of step 5e can be determined in a variety of ways. GBR-NMF provides two possible methods which include lack of progress and a maximum iteration count. Unless otherwise specified, the maximum iteration count is used since the lack of progress criterion depends greatly on the value of $\delta$ provided. Data sets with large variability may not be able to attain a small epsilon value. Contrarily, if the specified $\delta$ value is too large then the stopping criterion will be achieved too quickly, leaving room for improvement. To battle this, a high maximum iteration count of 50,000 is set to ensure a local maximum is reached.

4 Applications

4.1 Simulated Data

A simulation was generated in order to test the performance of the algorithm using $n = 400$ observations, $p = 2000$ variables, $g = 4$ groups, and $q = 7$ true underlying factors. This simulation was created to test the accuracy with respect to the
clustering aspect of the algorithm as well as how it does predicting scores on known factors. The simulation stipulates that each group differentiates itself from the others by having one unique factor that the others did not. Each one of the four groups had a unique factor, along with three other factors that were common with the other groups. Thus, when running the algorithm with respect to this data, the first four columns of the $W$ matrix are set to be the known classification matrix of the model. To add to the simulations, one of the three common features in $S$ were held as a constrained factor to determine how well its true scores were being found. The matrices below further illustrate which columns and rows in the model are being held constant (in bold) throughout the updating process.

$$W = \begin{pmatrix}
1 & 0 & 0 & w_{1,5} & w_{1,6} & w_{1,7} \\
0 & 1 & 0 & w_{2,5} & w_{2,6} & w_{2,7} \\
0 & 0 & 1 & w_{3,5} & w_{3,6} & w_{3,7} \\
0 & 0 & 0 & w_{4,5} & w_{4,6} & w_{4,7} \\
1 & 0 & 0 & w_{5,5} & w_{5,6} & w_{5,7} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 1 & w_{400,5} & w_{400,6} & w_{400,7}
\end{pmatrix} \quad (16)$$

$$A = \begin{pmatrix}
(a_{1,1} & 0 & 0 & 0 & 0 & 0 & 0) \\
0 & a_{2,2} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & a_{3,3} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & a_{4,4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & a_{5,5} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & a_{6,6} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & a_{7,7}
\end{pmatrix} \quad (17)$$

$$S = \begin{pmatrix}
s_{1,1} & s_{1,2} & \cdots & s_{1,2000} \\
s_{2,1} & s_{2,2} & \cdots & s_{2,2000} \\
s_{3,1} & s_{3,2} & \cdots & s_{3,2000} \\
s_{4,1} & s_{4,2} & \cdots & s_{4,2000} \\
s_{5,1} & s_{5,2} & \cdots & s_{5,2000} \\
s_{6,1} & s_{6,2} & \cdots & s_{6,2000} \\
s_{7,1} & s_{7,2} & \cdots & s_{7,2000}
\end{pmatrix} \quad (18)$$

Entries for $W$, $A$, and $S$ were all created using random uniform distributions and then scaled accordingly such that the column sums of $W$ are equal to $n$ and the area under the curve of the rows of $S$ were equal to 1. Thus, the diagonal elements of $A$ were scaled inversely such that the overall reconstructions would remain unchanged.

These data sets were created and decomposed via the GBR-NMF 100 different times in order to ensure the accuracy of the results. The residual sum of squares (RSS) comparing the true $S$ with the estimated $S$ was then calculated and a mean and standard error of these numbers were taken across all features (found in Table 1). This was similarly done with the resulting scores matrix $W$ and auxiliary matrix $A$ to quantify the algorithm’s overall estimation performance. More specifically, to
determine how close the values of the calculated matrices are to the true values. Table 1 displays that the algorithm does exceptionally well in predicting the features in both simulations. Meanwhile, the RSS of the scores matrix was also smaller in comparison with the RSS regarding the $A$ matrix. This is primarily due to the scalings on the two outside matrices which causes the values to become quite close. Therefore, a preferred comparison would be to the unconstrained model, which is also displayed in Table 1. This shows a drastic increase in both the accuracy of the scores and factors in comparison with the original model. This is due to the fact that we are holding constraints on the matrices $W$ and $S$. As we are specifying true groups, as well as a true factor, one may expect that the GBR-NMF model would perform better in this case. Therefore, we do not recommend GBR-NMF to be a direct replacement of the conventional NMF algorithm. It should instead be considered as an alternative method to be used when there is prior knowledge of the factor or group structure available.

It is important to note that this is a direct comparison between the standard NMF model and GBR-NMF. Comparisons to PCA or the semi-supervised NMF model of Liu et al. [25] would be inaccurate as a direct comparison is not available. Principal components acquired from PCA are orthogonal with respect to one another, making it difficult to make a direct comparison to the true factors created. Similarly, the different constraints in the work of Liu et al. [25] create different interpretations of the factor matrices.

### 4.2 Facial Expression Data

A common research application of nonnegative matrix factorization is data regarding facial expressions [38–40] typically for the identification of certain facial expressions or people. The pain expression data set [41] consists of 84 observations of $241 \times 181$ pixel facial images. These observations can be further separated into the faces of 12 different women; each of which displays 7 different facial expressions. An average face of each facial expression (aside from neutral) is displayed in row one of Fig. 1. These six facial expressions are the grouping variable of interest in this data set and are placed as

| W constraint held | Standard NMF | GBR-NMF |
|-------------------|--------------|---------|
|                   | Average      | Standard error |
| W RSS             | $4.7 \times 10^9$ | $1.9 \times 10^{10}$ |
| H RSS             | $4.0 \times 10^{-5}$ | $1.0 \times 10^{-5}$ |
|                   | Average      | Standard error |
| WA RSS            | $2.6 \times 10^9$ | $7.3 \times 10^8$ |
| S RSS             | $8.0 \times 10^{-6}$ | $4.0 \times 10^{-6}$ |
a grouping constraint in the $W$ matrix in the GBR-NMF. Along with the grouping constraints, twelve features were also constrained to be the neutral facial expressions of the different women. These constraints are determined to test what bases are acquired from the grouping variables as well as how close one can get to the true facial expression using a person’s neutral face and grouping information. From these neutral faces, the other bases are recovered and built on the top of this neutral face in order to recover the original images. The six bases recovered using GBR-NMF resemble key facial features explanatory of each emotion. Row two of Fig. 1 identifies the unconstrained features retrieved from the model. The fourth image of the second row displays a smile expression, which is portrayed well by highlighting the teeth and perked cheeks. These results are more interpretable and separated than those that are recovered from the standard NMF algorithm, as this algorithm tends to recover features that appear to more closely resemble particular women’s faces rather than emotion-specific representations. The standard NMF results are shown in row three of Fig. 1 which displays the six features of an unconstrained NMF sorted to best align with the original expressions.

This application displays that we are acquiring better separation in factors of the GBR-NMF model when compared to those of standard NMF. These interpretability
benefits could be assumed to result in a decrease in overall model performance. As we are using a similar update structure to that of standard NMF, one can only expect to match its reconstruction error at best. However, as demonstrated in Fig. 2, the reconstructions still do compare quite well to those of the unconstrained NMF model, making this a worthy trade-off.

4.2.1 Predicting an Unseen Facial Expression

In this section, we consider the possibility of generating previously unseen images using the GBR-NMF algorithm. Suppose one wanted to predict a person’s facial expression that is not included in the original model fit. Then the GBR-NMF could be used in order to predict what that expression might look like. The only information that would be needed in order to do so is the acquired bases from the model, as well as one’s neutral facial expression (i.e., the $\mathbf{S}$ matrix). This problem can be tackled in two different ways. The first scenario involves simply removing one observation in the data set and attempting to predict the face that was removed. The second requires omitting a person’s entire set of facial expressions from the model and predicting what their facial expression would look like using

Fig. 2 An original picture of a woman’s face (left) along with reconstructions from the GBR-NMF algorithm (center) and the standard NMF algorithm (right)

Fig. 3 True facial expression (left) in comparison with GBR-NMF prediction of single missing facial expression (center) and GBR-NMF missing facial expression prediction using only neutral face (right)
only the neutral facial expression and the bases acquired from the model. The predictions of these faces were made by scoring 1 on the desired facial expression and 1 on the neutral face, and using this linear combination as the prediction.

Acquisitions of the predicted faces, along with the true facial expression, are shown in Fig. 3. It is clear that these reconstructions are not as close to the original image as the results shown in Fig. 2. However, the reconstructions in Fig. 3 illustrate a predictive capability for the semi-supervised approach that is not achievable through the conventional unsupervised approach.

4.3 Raman Spectroscopic Cancer Data

In this example, Raman spectroscopy can be used to provide detailed information on the changes in relative concentrations of classes of molecular compounds (proteins, lipids, DNA, metabolites, etc.) within cellular and tissue environments exposed to ionizing radiation used in cancer therapy [10, 11, 42, 43]. Matthews et al. [11] introduce data containing 3240 spectra containing 582 Raman intensities (arbitrary units) at different wavenumbers (cm\(^{-1}\)). These spectra are separated into three different groups based on cell type, specifically lung (H460), breast (MCF7), and prostate (LNCaP) tumors. The spectra, overlaid with an average of each cell type, can be seen in the Fig. 4. In Raman spectroscopy, chemical constituents are identified through comparison with known chemical constituent spectra and literature line lists exist for the most common cellular components. Matthews et al. [11] performed a principal component analysis on these spectra and found that the cellular response to radiation exhibited a strong glycogen signal, identified as the 1st principal component in the data. The aim of this study is to further expand on the work of Matthews et al. [11] by recovering more biological bases using GBR-NMF. Using the GBR-NMF allows for the constraint of the factor matrix, ensuring the previously recovered glycogen is specified

![Fig. 4](image) Three averaged Raman spectra displaying three different cell lines which include H460 (red), MCF7 (blue), and LNCaP (green) tumor cells. Gray lines are the observed Raman spectra. Unit displayed on the y-axis is arbitrary intensity (AI).
in the model. Further recovered bases are then analyzed to discover new biological information with respect to the chemical composition of the cells in the dataset.

There are two main, and related, disadvantages to using PCA in this context. Firstly, permitting negative scores and component coefficients violates the physical understanding of the chemical composition of a cell — reconstructions of the spectra could be built in a subtractive manner where an overabundance of one chemical is estimated and then subtracted via a negative score on another chemical. Secondly, and perhaps related to the problems with previous assumptions, Matthews et al. [11] found a lack interpretability in additional principal components of the model. The standard NMF algorithm gains a key advantage over PCA due to the fact that it is forced to be an additive model with respect to chemical constituents, which match the physical construction of Raman spectroscopic data. This is demonstrated by Luce et al. [31] using the Successive Nonnegative Projection Algorithm to decompose time-resolved Raman spectroscopy into constituent spectra and kinetics of the underlying reactions. However, this method is only presented as an unsupervised learning technique and restricts one from being able to specify the previously suspected constituent spectra. Similarly, Deng et al. [44] perform unconstrained NMF on the data set of Matthews et al. [11] and not only discover glycogen, but also a spectrum very similar to that of a lipid. While using standard NMF is beneficial over PCA, the use of the GBR-NMF allows making use of the previous results from Matthews et al. [11] wherein an important chemical, glycogen, was discovered and biologically confirmed. It also ensures that the data will be decomposed into a linear combination of predominantly known features, which is something that cannot be guaranteed by the conventional approach. As such, GBR-NMF is utilized by constraining an acquired spectra of pure glycogen within S and further constraining W with the known cell lines in the data. We permit one further basis row and group column to be fully unconstrained.

![Fig. 5](image)

**Fig. 5** Scores on the five (four known/one unknown) chemical bases with respect to the three different cell lines (lung (red), breast (blue), prostate (green)). The Raman spectra in the bottom right corner are that of the unknown feature.
Figure 5 provides a scatterplot matrix of the scores from $W$ on the off diagonal and the recovered spectral bases along the diagonal. The first three bases along the diagonal were recovered under group constraints on the three tumor types. We note that all three bases in these cases closely approximate a standard cell spectra, with slight variations among some wavelengths. The glycogen scores for the green group (prostate, LNCaP) largely fall below a score of 1.5, which cannot be said about the other two cell lines. This further supports the result put forward by Matthews et al. [11] that LNCaP prostate tumors exhibit lower glycogen production in response to radiation than the other two cell lines (H460, MCF7). Furthermore, LNCaP prostate tumors are also known to be more sensitive to ionizing radiation than those of H460 and MCF7. Interestingly, the estimated results from GBR-NMF show that the prostate cells are not fully differentiable from the other cell lines based on glycogen alone, as there is substantial overlap in the glycogen scores, even if the LNCaP cells have lower, on average, scores. However, if we consider the fifth, fully unconstrained basis that was recovered (fifth diagonal along Fig. 5), we also see lower scores on average from the LNCaP cells. In fact, if the scores are viewed in tandem between glycogen and this currently unknown spectral component, we see complete separation between LNCaP cells (green) and the H460 and MCF7 cell lines (blue and red). Thus, applying the GBR-NMF method has provided a new spectral component to investigate with regard to potential chemicals that could be indicative of cell radiosensitivity. The current result presented here allows for a more direct focus on individual biochemical cellular response to radiotherapy as compared with the original basic PCA analysis.

One may be curious how this compares to the application of a standard NMF algorithm on these data. Figure 6 displays a scatter plot of the scores recovered when using standard NMF with $q = 5$. It should be noted that this plot is considerably different compared to that of Fig. 5. This is because this algorithm is completely unconstrained.

Fig. 6 Scores of the five unknown chemical bases recovered using a conventional NMF algorithm. The spectra of the unknown bases are displayed on the diagonal of the scatter plot.
Therefore, it does not allow us to constrain the bases or specify which spectra belong to which cell lines. As a result, these factors are completely unknown and each one needs to be carefully analyzed in order for specialists to match them to an appropriate chemical constituent. In fact, there is the possibility that none of the recovered factors resemble the chemical bases of interest. In the case of this particular data set, factors that have been identified include one which is similar to glycogen as well as a “lipid-like basis” [44]. However, it was already prior knowledge that glycogen is a key contributor in differentiating these spectra. As such, specifying the true glycogen spectra is more appropriate. More importantly, the specified constraints used in Fig. 5 allow us to identify an important relationship between glycogen and the single unknown basis and how this relationship differs with respect to prostate cancer. Further work is required to elucidate the biochemistry of radiation response and is beyond the scope of this manuscript.

A natural question that arises is how well the model estimated by GBR-NMF approximates the original spectra. Consider the plot in Fig. 7 showing two spectra, one being a randomly selected spectrum from the original data and the other being a reconstruction of that spectrum from GBR-NMF. From a spectroscopic viewpoint, this would be considered a suitable reconstruction.

5 Summary

A novel group and basis-restricted nonnegative matrix factorization (GBR-NMF) algorithm was developed to input known bases and/or known groups within a non-negative matrix factorization model. Updates were formulated and then tested via simulations and applications to real data: specifically, digital facial images and Raman spectroscopy. GBR-NMF was used to specify the known to exist a priori factors and groups in the models. Regarding the facial expression data, GBR-NMF displayed significant improvements on standard NMF with respect to the interpretability of the factors discovered under the group constraints. For the spectral data, the semi-supervised approach permitted by the GBR-NMF framework can aid the
scientific process by allowing pre-specification of known constituent chemicals and simultaneously suggesting potential constituents through the unconstrained factors.

Future methodological work could consider adjustments made to the auxiliary matrix such as adjusting the off-diagonal elements to be of non-zero value. Further exploration of $e$ in Eq. 5 should also be considered. Namely, the development of new updating procedures with respect to different data distributions and how the algorithm is affected when the assumption of Gaussian distributed error is violated. On the application side, the authors and collaborators are currently investigating the unknown Raman spectrum acquired (fifth diagonal of Fig. 5) from applying GBR-NMF to the data from Matthews et al. [11] for potential biological importance in indicating tumor radiosensitivity — such work has important ramifications for personalized radiotherapy programs. Research regarding error distributions of Raman spectroscopic data sets should also be investigated further, as the process of baseline subtraction and setting area under the curve of spectra to equal 1 could drastically affect this term.

### Appendix A: Derivation of Updating Algorithm

As specified in section 3.2, this does not account for the nonnegativity constraints. To do so, a Lagrange function $\mathcal{L}$ with multipliers $\alpha_{ij}$, $\beta_{ij}$, and $\gamma_{ij}$ is required in order to constrain $W$, $A$, and $S$ such that $w_{ij} \geq 0$, $a_{ij} \geq 0$ and $s_{ij} \geq 0$, respectively. This gives the following Lagrange function (with $\alpha = [\alpha_{ij}]$, $\beta = [\beta_{ij}]$, and $\gamma = [\gamma_{ij}]$)

$$\mathcal{L} = Tr(XX^T) - 2Tr(WASX^T) + Tr(WASS^TATW^T) + Tr(\alpha W^T) + Tr(\beta A^T) + Tr(\gamma S^T)$$

Now finding the derivatives of $\mathcal{L}$ with respect to $W$, $A$, and $S$, we have

$$D_F = \frac{1}{2} || X - WAS ||^2$$

$$\propto || X - WAS ||^2$$

$$= (\sqrt{Tr((X - WAS)(X - WAS)^T)})^2$$

$$= Tr((X - WAS)(X - WAS)^T)$$

$$= Tr((X - WAS)(X^T - S^TA^TW^T))$$

$$= Tr(XX^T - XS^TA^TW^T - WASX^T + WASS^TATW^T)$$

$$= Tr(XX^T) - Tr((XS^TA^TW^T)^T) - Tr(WASX^T) + Tr(WASS^TATW^T)$$

$$= Tr(XX^T) - Tr((XS^TA^TW^T)^T) - Tr(WASX^T) + Tr(WASS^TATW^T)$$

$$= Tr(XX^T) - 2Tr(WASX^T) + Tr(WASS^TATW^T)$$
\[
\frac{\partial L}{\partial W} = \frac{\partial}{\partial W}(Tr(XX^T) - 2Tr(WASX^T) + Tr(WASS^T A^T W^T) \\
+ Tr(\alpha W^T) + Tr(\beta A^T) + Tr(\gamma S^T)) \\
= \frac{\partial}{\partial W} Tr(XX^T) - 2 \frac{\partial}{\partial W} Tr(WASX^T) + \frac{\partial}{\partial W} Tr(WASS^T A^T W^T) \\
+ \frac{\partial}{\partial W} Tr(\alpha W^T) + \frac{\partial}{\partial W} Tr(\beta A^T) + \frac{\partial}{\partial W} Tr(\gamma S^T) \\
= -2 \frac{\partial}{\partial W} Tr(WASX^T) + \frac{\partial}{\partial W} Tr(WASS^T A^T W^T) + \frac{\partial}{\partial W} Tr(\alpha W^T) \\
= -2X^TAS + 2WASS^TA^T + \alpha \\
= 0
\]

\[
\frac{\partial L}{\partial A} = \frac{\partial}{\partial A}(Tr(XX^T) - 2Tr(WASX^T) + Tr(WASS^T A^T W^T) \\
+ Tr(\alpha W^T) + Tr(\beta A^T) + Tr(\gamma S^T)) \\
= \frac{\partial}{\partial A} Tr(XX^T) - 2 \frac{\partial}{\partial A} Tr(WASX^T) + \frac{\partial}{\partial A} Tr(WASS^T A^T W^T) \\
+ \frac{\partial}{\partial A} Tr(\alpha W^T) + \frac{\partial}{\partial A} Tr(\beta A^T) + \frac{\partial}{\partial A} Tr(\gamma S^T) \\
= -2 \frac{\partial}{\partial A} Tr(WASX^T) + \frac{\partial}{\partial A} Tr(WASS^T A^T W^T) \frac{\partial}{\partial A} Tr(\beta A^T) \\
= -2W^TXS + 2W^T WASS^T + \beta \\
= 0
\]

\[
\frac{\partial L}{\partial S} = \frac{\partial}{\partial S}(Tr(XX^T) - 2Tr(WASX^T) + Tr(WASS^T A^T W^T) \\
+ Tr(\alpha W^T) + Tr(\beta A^T) + Tr(\gamma S^T)) \\
= \frac{\partial}{\partial S} Tr(XX^T) - 2 \frac{\partial}{\partial S} Tr(WASX^T) + \frac{\partial}{\partial S} Tr(WASS^T A^T W^T) \\
+ \frac{\partial}{\partial S} Tr(\alpha W^T) + \frac{\partial}{\partial S} Tr(\beta A^T) + \frac{\partial}{\partial S} Tr(\gamma S^T) \\
= -2 \frac{\partial}{\partial S} Tr(WASX^T) + \frac{\partial}{\partial S} Tr(WASS^T A^T W^T) + \frac{\partial}{\partial S} Tr(\gamma S^T) \\
= -2A^T W^T X + 2A^T W^T WASS + \gamma \\
= 0
\]

Each of these is set to equal zero as the goal is to find a local optimum of \( D_F \) since it is unfeasible to find the global minimum. Applying the Karush-Kuhn-Tucker conditions, \( \alpha_{ij} w_{ij} = 0, \beta_j a_{ij} = 0, \) and \( \gamma_{ij} s_{ij} = 0 \) to the three equations results in the following
which leads to the following updating rules

\[
\begin{align*}
    w_{ij} & \leftarrow w_{ij} \frac{(XS^TA^T)_{ij}}{(W^TAX^T)_{ij}} \\
a_{ij} & \leftarrow a_{ij} \frac{(W^TXS^T)_{ij}}{(W^TWASS^T)_{ij}} \\
s_{ij} & \leftarrow s_{ij} \frac{(ATW^TX)_{ij}}{(ATW^TWAS)_{ij}}
\end{align*}
\] (20)

\[
(XS^TA^T)_{ij}w_{ij} - (W^TAW^T)_{ij}w_{ij} = 0
\]

\[
(W^TXS^T)_{ij}a_{ij} - (W^TWASS^T)_{ij}a_{ij} = 0
\] (19)

\[
(ATW^TX)_{ij}s_{ij} - (ATW^TWAS)_{ij}s_{ij} = 0
\]

### Appendix B: Proof of Algorithm Convergence

Appendix B proves Lemmas 3.3 and 3.4 in order to fulfill the proof requirements of Sect. 3.3.

#### B.1 Proof of Lemma 3.3

**Proof** It is clear that \( G(a, a) = F_{a_j}(a) \). According to the definition of an auxiliary function, it only needs to be shown that \( G(a, a') \geq F_{a_j}(a) \). This can be done using the Taylor series expansion of \( F_{a_j} \):

\[
F_{a_j}(a) = F_{a_j}(a^j) + F'_{a_j}(a - a^j) + \frac{1}{2} F''_{a_j}(a - a^j)^2
\]

with \( F''_{a_j} \) being the second-order derivative of \( F_{a_j} \). We can show that

\[
F_{a_j} = \left( \frac{\partial D}{\partial A} \right)_{ij} = (-2W^TXS^T + 2W^TWASS^T)_{ij}
\]

\[
F''_{a_j} = \left( \frac{\partial^2 D}{\partial A^2} \right)_{ij} = 2(W^TW)_{ij}(SS^T)_{ij}
\]

It is now equivalent to prove

\[
\frac{(W^TWASS^T)_{ij}}{a^j_{ij}} \geq \frac{1}{2} F''_{ij} = (W^TW)_{ij}(SS^T)_{ij}
\] (21)
Now we have
\[
(W^T W S S^T)_{ij} = \sum_{\ell=1}^{q} (W^T W A)_{ij} (S S^T)_{\ell j}
\] (22)
\[
\geq (W^T W A)_{ij} (S S^T)_{jj}
\] (23)
\[
\geq \sum_{\ell=1}^{q} (W^T W)_{ij} a_{\ell n} (S S^T)_{nn}
\] (24)
\[
\geq a_{ij} (W^T W)_{ij} (S S^T)_{jj}
\] (25)
which when rearranged proves Eq. 21. □

B.2 Proof of Lemma 3.4

**Proof** It is clear that $G(s, s) = F_{s_j}(s)$. According to the definition of an auxiliary function, it only needs to be shown that $G(s, s^t) \geq F_{s_j}(s)$. This can be done using the Taylor series expansion of $F_{s_j}$:

\[
F_{s_j}(s) = F_{s_j}(s^t_j) + F'_{s_j}(s - s^t_j) + \frac{1}{2} F''_{s_j}(s - s^t_j)^2
\]
with $F''_{s_j}$ being the second-order derivative of $F_{s_j}$. We can show that

\[
F_{s_j} = \left( \frac{\partial D}{\partial S} \right)_{ij} = (-2A^T W^T X + 2A^T W^T W A)_{ij}
\]
\[
F''_{s_j} = \left( \frac{\partial^2 D}{\partial S^2} \right)_{ij} = 2(A^T W^T W A)_{ii}
\]

It is now equivalent to prove
\[
\frac{(A^T W^T W A)_{ij}}{s^t_{ij}} \geq \frac{1}{2} F''_{ij} = (A^T W^T W A)_{ii}
\] (26)

. Now we have
\[
((A^T W^T W A)_{ij} = \sum_{\ell=1}^{q} (A^T W^T W A)_{ij} (S S^T)_{\ell j}
\] (27)
\[
\geq (A^T W^T W A)_{ij} s^t_{ij}
\] (28)
\[ s'_{ij} = s_{ij} \left( A^T W^T W A S \right)_{ii} \] (29)

which when rearranged proves Eq. 26.

Acknowledgements This research was supported by the Natural Sciences and Engineering Research Council of Canada through their Discovery Grants program (RGPIN-2014-06730 - Andrews, RGPIN-2020-07232 - Jirasek) and the Cluster of Research Excellence in Medical Physics and Data Analytics (PI: Jirasek) enabled via the University of British Columbia’s Eminence Fund. Infrastructure support was provided by the Canada Foundation for Innovation through their John R. Evans Leaders Fund program (#35578 - Andrews, #32859 - Jirasek PI).

Funding This research was supported by the Natural Sciences and Engineering Research Council of Canada through their Discovery Grants program (RGPIN-2014-06730 - Andrews, RGPIN-2020-07232 - Jirasek) and the Cluster of Research Excellence in Medical Physics and Data Analytics (PI: Jirasek) enabled via the University of British Columbia’s Eminence Fund. Infrastructure support was provided by the Canada Foundation for Innovation through their John R. Evans Leaders Fund program (#35578 - Andrews, #32859 - Jirasek PI).

Data Availability The simulation data is available on GitHub repository. The pain expression data set is publicly available [41] and the Raman spectroscopic data set can be made available upon reasonable request.

Code Availability Code for the GBR-NMF algorithm is available on GitHub repository. [45]

Declarations

Conflict of interest There is no conflict of interest to declare.

References

1. Jolliffe Ian (2011) Principal component analysis. Springer, New York
2. Gray Robert M (1990) Vector quantization. Read Speech Recognit 1(2):75–100
3. Harman Harry H (1976) Modern factor analysis. University of Chicago press
4. Lee Daniel D, Sebastian Seung H (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788
5. Wang Yu-Xiong, Zhang Yu-Jin (2012) Nonnegative matrix factorization: A comprehensive review. IEEE Trans Knowl Data Eng 25(6):1336–1353
6. Butler Holly J, Ashton Lorna, Bird Benjamin, Cinque Gianfelice, Curtis Kelly, Dorney Jennifer, Esmonde-White Karen, Fullwood Nigel J, Gardner Benjamin, Martin-Hirsch Pierre L et al (2016) Using raman spectroscopy to characterize biological materials. Nat Protoc 11(4):664
7. Feng Xu, Moy Austin J, Nguyen Hieu TM, Zhang Jason, Fox Matthew C, Sebastian Katherine R, Reichenberg Jason S, Markey Mia K, Tunnell James W (2017) Raman active components of skin cancer. Biomed Opt Exp 8(6):2835–2850
8. Pence Isaac, Mahadevan-Jansen Anita (2016) Clinical instrumentation and applications of raman spectroscopy. Chem Soc Rev 45(7):1958–1979
9. Pieters Sigrid, Heyden Yvan Vander, Roger Jean-Michel, D’Hondt Matthias, Hansen Laurent, Palagos Bernard, De Spiegeleer Bart, Remon Jean-Paul, Vervaet Chris, De Beer Thomas (2013) Raman spectroscopy and multivariate analysis for the rapid discrimination between native-like and non-native states in freeze-dried protein formulations. Eur J Pharm Biopharm 85(2):263–271
10. Harder Samantha J, Matthews Quinn, Isabelle Martin, Brolo Alexandre G, Lum Julian J, Jirasek Andrew (2015) A raman spectroscopic study of cell response to clinical doses of ionizing radiation. Appl Spectrosc 69(2):193–204
11. Matthews Quinn, Isabelle Martin, Harder Samantha J, Smazynski Julian, Beckham Wayne, Brolo Alexandre G, Jirasek Andrew, Lum Julian J (2015) Radiation-induced glycogen accumulation detected by single cell raman spectroscopy is associated with radioresistance that can be reversed by metformin. PloS ONE 10(8):e0135356
12. Koch H, Polepil S, Eisen K, Will S (2017) Raman microspectroscopy and multivariate data analysis: optical differentiation of aqueous d-and l-tryptophan solutions. Phys Chem Chem Phys 19(45):30533–30539
13. Lee Daniel D, Sebastian Seung H (2001) Algorithms for non-negative matrix factorization. Adv Neural Inf Process Syst 13:556–562
14. Févotte Cédric, Cemgil A Taylan (2009) Nonnegative matrix factorizations as probabilistic inference in composite models. In: 2009 17th European Signal Processing Conference, pp. 1913–1917. IEEE
15. Vavasis Stephen A (2010) On the complexity of nonnegative matrix factorization. SIAM J Optim 20(3):1364–1377
16. Donoho David, Stodden Victoria (2003) When does non-negative matrix factorization give a correct decomposition into parts? In: Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 1141–1148. NeurIPS
17. Lin Chih-Jen (2007) Projected gradient methods for nonnegative matrix factorization. Neural Comput 19(10):2756–2779
18. Zdunek Rafal, Cichocki Andrzej (2007) Nonnegative matrix factorization with constrained second-order optimization. Signal Process 87(8):1904–1916
19. Guan Naiyang, Tao Dacheng, Luo Zhigang, Yuan Bo (2012) Nenmf: an optimal gradient method for nonnegative matrix factorization. IEEE Trans Signal Process 60(6):2882–2898
20. Devarajan Karthik (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol 4(7):e1000029
21. Ding Chris, He Xiaofeng, Simon Horst D (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proceedings of the 2005 SIAM international conference on data mining, pp. 606–610. SIAM
22. Hartigan John A, Wong Manchek A (1979) Algorithm as 136: a k-means clustering algorithm. J Royal Stat Soc. Series C 28(1):100–108
23. Li Tao, Ding Chris, Jordan Michael I (2007) Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: 7th IEEE International Conference on Data Mining (ICDM 2007), pp. 577–582. IEEE
24. Wagstaff Kiri, Cardie Claire, Rogers Seth, Schrödl Stefan et al (2001) Constrained k-means clustering with background knowledge. In Icml 1:577–584
25. Liu Haifeng, Zhaohui Wu, Li Xuelong, Cai Deng, Huang Thomas S (2011) Constrained non-negative matrix factorization for image representation. IEEE Trans Pattern Anal Machine Intell 34(7):1299–1311
26. Zuyuan Yang Yu, Zhang Yong Xiang, Yan Wei, Xie Shengli (2018) Non-negative matrix factorization with dual constraints for image clustering. IEEE Trans Syst, Man, and Cybern: Syst 50(7):2524–2533
27. Arora Sanjeev, Ge Rong, Kannan Ravi, Moitra Ankur (2016) Computing a nonnegative matrix factorization—provably. SIAM J Comput 45(4):1582–1611
28. Gillis Nicolas (2014) Successive nonnegative projection algorithm for robust nonnegative blind source separation. SIAM J Imaging Sci 7(2):1420–1450
29. Gillis Nicolas, Luce Robert (2014) Robust near-separable nonnegative matrix factorization using linear optimization. J Machine Learn Res 15(1):1249–1280
30. Qing Qu, Nasrabadi Nasser M, Tran Trac D (2015) Subspace vertex pursuit: a fast and robust near-separable nonnegative matrix factorization method for hyperspectral unmixing. IEEE J Select Topics Signal Process 9(6):1142–1155
31. Luce Robert, Hildebrandt Peter, Kuhlmann Uwe, Liesen Jörg (2016) Using separable nonnegative matrix factorization techniques for the analysis of time-resolved raman spectra. Appl Spectrosc 70(9):1464–1475
32. Ding Chris, Li Tao, Peng Wei, Park Haesun (2006) Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 126–135.
33. Chen Gang, Wang Fei, Zhang Changshui (2009) Collaborative filtering using orthogonal nonnegative matrix tri-factorization. Inf Process Manag 45(3):368–379
34. Li Tao, Zhang Yi, Sindhwani Vikas (2009) A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 244–252
35. Pauca V Paul, Piper Jon, Plemmons Robert J (2006) Nonnegative matrix factorization for spectral data analysis. Linear Algebra Appl 416(1):29–47
36. Dempster Arthur P, Laird Nan M, Rubin Donald B (1977) Maximum likelihood from incomplete data via the em algorithm. J Royal Stat Soc: Series B 39(1):1–22
37. Jeff Wu CF (1983) On the convergence properties of the em algorithm. Annals of Stat 11(1):95–103
38. Buciu Ioan, Pitas, Ioannis (2004) Application of non-negative and local non negative matrix factorization to facial expression recognition. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., volume 1, pp. 288–291. IEEE
39. Guillaumet David, Vitria, Jordi (2002) Classifying faces with nonnegative matrix factorization. In Proc. 5th Catalan conference for artificial intelligence, pp. 24–31
40. Zhi Ruicong, Flierl Markus, Ruan Qiuqi, Kleijn W Bastiaan (2010) Graph-preserving sparse non-negative matrix factorization with application to facial expression recognition. IEEE Trans Syst, Man, and Cybern, Part B 41(1):38–52
41. Psychological Image Collection at Sterling. 2D face sets - pain expressions data at pics.stir.ac.uk. pics.stir.ac.uk, 2019. Accessed: 2019-09-12
42. Matthews Quinn, Jirasek Andrew, Lum Julian, Duan Xiaobo, Brolo Alexandre G (2010) Variability in raman spectra of single human tumor cells cultured in vitro: correlation with cell cycle and culture confluency. Applied Spectrosc 64(8):871–887
43. Paidi Santosh K, Diaz Paola Monterroso, Dadgar Sina, Jenkins Samir V, Quick Charles M, Grifﬁn Robert J, Dings Ruud PM, Rajaram Narasimhan, Barman Ishan (2019) Label-free raman spectroscopy reveals signatures of radiation resistance in the tumor microenvironment. Cancer Res 79(8):2054–2064
44. Deng Xinchen, Ali-Adeeb Ramie, Andrews Jeffrey L, Shreeves Phillip, Lum Julian J, Brolo Alexandre, Jirasek Andrew (2020) Monitor ionizing radiation-induced cellular responses with raman spectroscopy, non-negative matrix factorization, and non-negative least squares. Appl Spectrosc 74(6):701–711
45. Phillip Shreeves and Jeffrey L. Andrews. Group and basis restricted nonnegative matrix factorization. https://github.com/its-likeli-jeff/GBRNMF, 2020

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.