Dimensionality Reduction Techniques with Applications in Raman Spectroscopy

by

Phillip Shreeves

B.Sc. Hons., The University of British Columbia, 2018
M.Sc., The University of British Columbia, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The College of Graduate Studies

(Mathematics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

August, 2020

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis/dissertation entitled:

Dimensionality Reduction Techniques
with Applications in Raman Spectroscopy

submitted by Phillip Shreeves in partial fulfilment of the requirements of the degree of Master of Science

Dr. Jeffrey L. Andrews, I. K. Barber School of Arts & Sciences
Supervisor

Dr. Andrew Jirasek, I. K. Barber School of Arts & Sciences
Supervisory Committee Member

Dr. W. John Braun, I. K. Barber School of Arts & Sciences
Supervisory Committee Member

Dr. Sepideh Pakpour, School of Engineering
University Examiner

Abstract

Raman spectroscopy (RS) is an optical interrogation method used to identify molecules through inelastic light scattering. This process often results in the acquired data having high dimensionality, as each observation has multiple intensities over hundreds of wave numbers. As a result, a common progression in the analysis of the spectra is to first reduce the dimensionality of the data before further examination. This thesis details the use of different dimensionality reduction techniques, including an extension of nonnegative matrix factorization (NMF). NMF is the process of decomposing the data matrix of interest into two lower rank matrices, with the constraint that all matrices must be nonnegative. This algorithm is extended in order to impose constraints on the model by decomposing the one of the specified matrices into two further nonnegative matrices, creating three in total. The application of this technique handles the previously discussed issues while also having added interpretability in comparison to principal component analysis (PCA), which is currently used in RS.

Lay Summary

In this thesis, a novel algorithm, Group and Basis Restricted Nonnegative Matrix Factorization (GBR-NMF), is developed with motivations from Raman spectroscopy data sets. Raman spectroscopy is a field in medical physics that can identify molecules through an light scattering. These molecules are then decomposed, using GBR-NMF, into bases that make up said molecules. The GBR-NMF differs from standard NMF due to the fact that suspected bases and groups can be specified in the model. This model is compared to standard NMF on simulated data, image data, and the motivating data sets to demonstrate the capability of the GBR-NMF algorithm.

Contents

List of Tables

List of Figures

Acknowledgements

My deepest gratitude goes to Dr. Jeffrey Andrews for taking me on as his student and providing me with guidance over the past 3 years of my academic career. I would also like to thank those in the Raman spectroscopy group, led by Dr. Andrew Jirasek, for accepting me in their group and educating me in their field. I would like to thank Dr. John Braun, Dr. Jirasek, and Dr. Andrews for their contributions on my advisory committee. Finally, I would like to thank both my family and friends for their endless support.

Chapter 1

Introduction

As the size of data sets are increasing, the world of data analytics is also expanding. Largely due to the high number of variables, this results in even the most commonly known methods to be infeasible at times. This issue has resulted in an increased focus on a form of data transformation, known as dimensionality reduction, to arise. 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, Jolliffe, 2011), vector quantization (VQ, Gray, 1990), factor analysis (Harman, 1960), and nonnegative matrix factorization (NMF, Lee and Seung, 1999). All of these techniques follow two key properties of dimensionality reduction specified by Wang and Zhang (2012), which includes reducing the dimensionality of the data and recovering factors 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 are a much more parts-based representation, rather than holistic, which can be beneficial in many applications. For example, Lee and Seung (1999) first introduce NMF by applying it to a data set containing facial images. NMF was 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 said parts. The standard NMF algorithm is an unsupervised learning technique and is unfortunately sometimes unable to tease out known to exist a priori factors. 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 modelling 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, multiplexed, information on a range of

molecular constituents within a single sample acquisition (Butler et al., 2016; Feng et al., 2017; Pence and Mahadevan-Jansen, 2016). 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. 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 uncorrelated. This is because the increase of one chemical in a spectrum typically constitutes having less than another and vice versa. Thus other dimensionality reduction techniques, such as NMF, are more appropriate for application to these data sets. In this thesis, 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 said data in order to further expand knowledge in the field of RS.

1.1 Objective

Raman spectroscopy techniques typically use principal component analysis in order to reduce the dimensionality of the data. However, these data sets violate some of the assumptions one needs to make. This includes the assumption that components can take on negative values and that components are orthogonal of one another. As a result, this is an inappropriate model to use and the data is much better suited for nonnegative matrix factorization. Using NMF adds the constraint that all values must be nonnegative and also removes the assumption of uncorrelated components. In this thesis, an extension of nonnegative factorization is proposed which allows for previously expected factors as well as known groups to be specified in the model, increasing overall interpretability in comparison to conventional NMF.

1.2 Thesis Organization

1.2.1 Chapter 2

The second chapter reviews modern techniques that will be used and expanded upon throughout the thesis. Methods such as principal component analysis, the least absolute shrinkage and selection operator, logistic regression, and nonnegative matrix factorization are discussed in detail.

1.2.2 Chapter 3

Chapter three expands upon the nonnegative matrix factorization section discussed in chapter two. It does so by providing a novel extension of the algorithm through added constraints, allowing for both clustering applications as well as the ability to specify suspected factors in the model. These imposed constraints, along with the algorithm's updating process and optimization criterion, are outlined in this section.

1.2.3 Chapter 4

This chapter provides proof of concept of the algorithm provided in chapter three. It shows applications of the algorithm on multiple data sets and also displays other dimensionality reduction techniques that can be applied in Raman spectroscopy examples.

1.2.4 Chapter 5

The final chapter of the thesis concludes and summarizes the work that was completed throughout. The impact that this thesis provides and future work to be performed are also discussed.

Chapter 2

Background

Herein, several common and modern techniques for dimensionality reduction and modelling are reviewed.

2.1 Principal Component Analysis

One of the most commonly used dimensionality reduction techniques, Principal Component Analysis (PCA) was first developed by Pearson (1901) and is used in many different situations and some primary goals of the model include variable selection, classification, and dimensionality reduction (Wold et al., 1987). PCA uses linear combinations of the original variables in the data set to generate new variables. The new variables created are uncorrelated and are made to retain as much variance from the original data as possible, while also reducing dimensionality (Jolliffe and Cadima, 2016). This by calculating the first principal component (PC) by achieving the maximum possible variance, then calculating the second PC in the same fashion while ensuring it is uncorrelated with the first (Jolliffe, 1990). The process of calculating components continues until the number of PCs matches the total number of variables in the original data set. In particular, the weights of the linear combination for each of the new variables (PCs) are found via the eigenvectors of the covariance/correlation matrix of the original variables. From there, the number of PCs to be kept can be determined in regards to how much variance one desires to keep. The eigenvalues of said eigenvectors are then proportional to the percentage of variance from the original data associated with that PC. Each observation is then mapped to the PC space via weights on each original variable. Notably both the eigenvectors and weights can take on negative values. While removing correlated features is not desirable in certain applications, it often is a benefit of PCA. It is also advantageous to be able to choose how much information one wants to retain with respect to choosing a number of principal components. The caveat regarding choosing said number is that information is guaranteed to be lost unless all components are retained. Along with loss of information, PCA also becomes difficult to interpret compared to the original variables in complex

applications. For more comprehensive information on Principal Component Analysis, see Jolliffe (1986), Abdi and Williams (2010), and Jolliffe (2011).

2.2 Least Absolute Shrinkage and Selection Operator

The "Least Absolute Shrinkage and Selection Operator"(Tibshirani, 1996), more commonly referred to as LASSO, is a well known technique used for variable selection and shrinkage of coefficients in generalized linear models. First developed with respect to the commonly known ordinary least-squares linear regression model, this method has expanded to other familiar models which includes the Cox regression (Tibshirani, 1997) and the logistic regression (Shevade and Keerthi, 2003). In an application of the ordinary least squares regression, coefficients are estimated by minimizing the following function

$$(\alpha, \boldsymbol{\beta}) = \underset{\alpha,\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j=1}^{p} (\beta_j x_{ij}))^2. \tag{2.1}$$

However, commonly known issues with said algorithm include the large amounts of variability between estimated variables and true values and lack of interpretability with respect to high dimensionality models (Tibshirani, 1996). These issues can be resolved through the addition of a penalty term, which is added to the objective function as shown below

$$(\alpha, \boldsymbol{\beta}) = \underset{\alpha,\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j=1}^{p} (\beta_j x_{ij}))^2 + \lambda \mid\mid \boldsymbol{\beta} \mid\mid_1 \tag{2.2}$$

where α and $\boldsymbol{\beta}$ represent the intercept and model coefficients respectively. The additional term is effectively equivalent to minimizing Equation 2.1 using a constraint ($\sum \mid \beta_j \mid < s$) on the size of the coefficients in the model (Tibshirani, 2011). This, in turn, allows for the adjustment of said λ term in order to control the total amount of shrinkage of the overall model. For example, as λ approaches infinity, it is notable that the beta terms would need to be shrunken to zero in order to optimize the function. Contrarily, if λ takes on a very small value, then the coefficients are not substantially restricted in magnitude which may result in large values. Additional information regarding LASSO can be found in sources by Kwon et al. (2013) and Hastie et al. (2015).

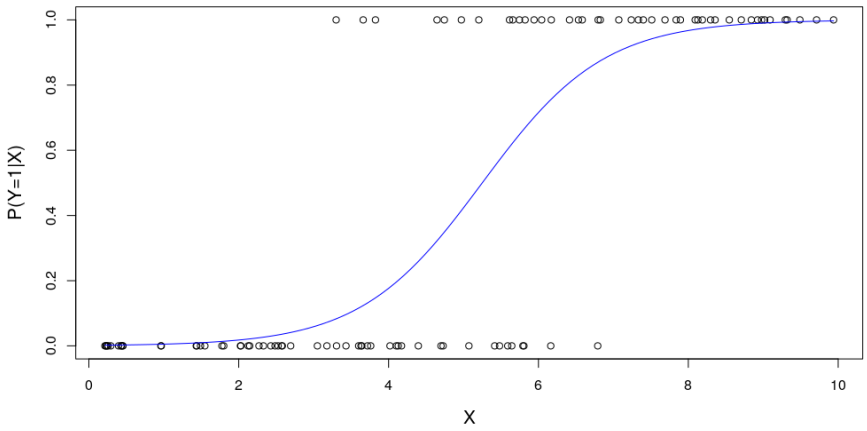

Figure 2.1: Univariate example of logistic regression with the response variable being observation class.

2.3 Logistic Regression

Logistic regression is a classification technique used to predict an observation's class with the variables at hand and determine the relationship between them. The model in a two group scenario is characterized using the following equation

$$P(Y = 1 \mid \mathbf{x}_i) = \frac{1}{1 + e^{-(\alpha + \mathbf{x'}_i \boldsymbol{\beta})}} \tag{2.3}$$

where $P(Y = 1 \mid \mathbf{x}_i)$ represents the probability of the observation belonging to the class of interest. This causes said model to be frequently used in the field of epidemiology due to $P(Y = 1 \mid \mathbf{x}_i)$ taking a probabilistic value. In univariate cases, this model is well known to have an S-shaped probability curve as shown in Figure 2.1 which allows for a distinct cut-off point between classes. This technique can further be expanded to a multi-class scenario (Böhning, 1992) that takes similar form to its binomial counterpart. The probability of observation i belonging to group $g \in \{1, ..., G\}$ can be calculated in the subsequent manner

$$P(Y = g \mid \mathbf{x}_i) = \frac{e^{\alpha + \mathbf{x}_i' \boldsymbol{\beta}^{(g)}}}{1 + e^{\alpha + \mathbf{x}_i' \boldsymbol{\beta}^{(g)}}} \text{ for } g = 1, ..., G - 1 \tag{2.4}$$

$$P(Y = G \mid \mathbf{x}_i) = \frac{1}{1 + e^{\alpha + \mathbf{x}_i' \boldsymbol{\beta}^{(G)}}} \tag{2.5}$$

where $\beta_j^{(g)}$ represents the j^{th} coefficient in the g^{th} class model and the G^{th} group represents the reference variable. Further information on this topic can be found in sources by Kleinbaum et al. (2002), Hilbe (2009), and Hosmer Jr et al. (2013).

2.3.1 LASSO Logistic Regression

Unlike the conventional LASSO regression, LASSO logistic regression imposes a constraint on the negative log-likelihood of the model. This likelihood is given by Huang et al. (2008) as

$$\ell_n(\beta) = -\sum_{i=1}^{n}(Y_i \log \pi(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - Y_i)\log[1 - \pi(\mathbf{x}_i'\boldsymbol{\beta})]) \tag{2.6}$$

where $\pi(t) = e^t/(1 + e^t)$ and is constrained such that

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min(n^{-1}\ell_n(\boldsymbol{\beta}) + \lambda \mid\mid \boldsymbol{\beta} \mid\mid_1) \tag{2.7}$$

by Fonti and Belitser (2017). This constraint is added in a similar fashion when LASSO is applied to the multinomial logistic regression (Zhai, 2018).

2.4 Nonnegative Matrix Factorization

Nonnegative matrix factorization was originally developed by Lee and Seung (1999) as a method of finding the major underlying factors of a model and seeing how each individual observation is composed of the 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

$$\mathbf{X} \approx \mathbf{WH} \tag{2.8}$$

where with q underlying factors, we can describe the span of the columns of $\mathbf{H}$ as the q-dimensional subspace and $\mathbf{W}$ as the scores for each observation on that 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 $\mathbf{X}$

$$\mathbf{X} = \mathbf{WH} + \varepsilon \tag{2.9}$$

with $\varepsilon \in \mathbb{R}^{n \times p}$ represents the residual or noise matrix of the model. This noise $\varepsilon = \mathbf{X} - \mathbf{WH}$, is a key component in the model as its assumed probability distribution assists in determining what objective function the data needs to be optimized with respect to (Wang and Zhang, 2012). For example, assuming a Gaussian distributed residue matrix, the Frobenius norm

$$D_F(\mathbf{X} \mid \mathbf{WH}) = \frac{1}{2} \mid\mid \mathbf{X} - \mathbf{WH} \mid\mid_F^2 = \frac{1}{2} \sum_{ij} (x_{ij} - [\mathbf{WH}]_{ij})^2 \tag{2.10}$$

is the function that is to be optimized (Lee and Seung, 2001; Wang and Zhang, 2012). Another function commonly used is the generalized Kullback-Leibler divergence (GKLD) which is used when assuming Poisson distributed noise. For this thesis, we will assume Gaussian noise as the simulated data sets were created with said noise and preliminary diagnostics regarding real data sets for complex Raman spectra such as those in Sections 4.3 and 4.3.1 did not reject the assumption of Gaussian noise.

Lee and Seung (2001) describe that NMF can be formulated as two optimization problems, which are to minimize both $\mid\mid \mathbf{X} - \mathbf{WH} \mid\mid$ and $D(\mathbf{X} \mid \mathbf{WH})$ simultaneously and while these two functions are convex in $\mathbf{W}$ or $\mathbf{H}$ individually, they are unfortunately not convex together. Because of this, finding global minima is not guaranteed and algorithms which find local minima are predominantly considered at the moment. The prototypical algorithm for finding these matrices is an alternating multiplicative updating process that uses the following updates:

$$h_{ij} \leftarrow h_{ij} \frac{(\mathbf{W}^T\mathbf{X})_{ij}}{(\mathbf{W}^T\mathbf{WH})_{ij}}, \qquad w_{ij} \leftarrow w_{ij} \frac{(\mathbf{XH}^T)_{ij}}{(\mathbf{WHH}^T)_{ij}}. \tag{2.11}$$

This updating process is considered slow by most accounts and scales dramatically in elapsed time when requiring both the rank and dimension of $\mathbf{X}$ to increase. Suggestions to increase the performance of the algorithm include applying gradient descent algorithms (Lin, 2007) or use of the conjugate gradient (Zdunek and Cichocki, 2007). While NMF is advantageous due to it's added interpretability in comparison to PCA, it does fall victim to slower computation speed (Devarajan, 2008). Aside from the assumption of nonnegativity, Lee and Seung (1999) also note that no further assumptions are made about the statistical independencies of the underlying factors.

2.4.1 Constrained Nonnegative Matrix Factorization

Various authors have considered constraints to address questions concerning clustering and classification with the NMF algorithm. Ding et al. (2005) 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 (Hartigan and Wong, 1979) 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. (2007) in a semi-supervised clustering manner, where limited knowledge of cluster membership is present. Wagstaff et al. (2001) specify how semi-supervised clustering is an algorithm performed with respect to two separate constraints. These two constraints are that there is the existence of must-link observations such that said observations must be grouped into the same cluster and that theres 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. (2011) address these constraints by introducing the decomposition of the feature scores matrix $\mathbf{W}$ 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 $\mathbf{C} = [\, \mathbf{c}_1 \; \mathbf{c}_2 \; \cdots \; \mathbf{c}_g \,]$ 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 $\mathbf{W} = \left[\begin{smallmatrix} \mathbf{C} & 0 \\ 0 & \mathbf{I} \end{smallmatrix}\right]$. Both $\mathbf{W}$ 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 is able to provide a contribution to the updating process described in Section 3.2. Yang et al. (2018) further expand upon the work of Liu et al. (2011) by adding in a sparsity constraint, making it a dual constrained algorithm.

Chapter 3

Methodology

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

The conventional NMF algorithm proposed by Lee and Seung (1999) is a completely unsupervised algorithm which is cause for concern in certain applications sometimes. This is because 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. (2011) by similarly utilizing a further decomposition of the data of interest into matrices $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{S}$ such that

$$\mathbf{X} = \mathbf{WAS} + \epsilon, \tag{3.1}$$

where it is apparent that $\mathbf{H}$ from Equation 2.8 is further decomposed into a $q \times q$ auxiliary matrix $\mathbf{A}$ and the $q \times p$ matrix $\mathbf{S}$ containing the underlying factors. However, this algorithm differs in the way that restrictions are imposed on the matrices. Liu et al. (2011) restrict their $\mathbf{W}$ matrix as stated in Section 2.4.1, whereas the new algorithm restricts a number of columns in $\mathbf{W}$ through the use of a complete indicator matrix and also restricts some rows in $\mathbf{S}$. The added restrictions allow for one to specify known groups in $\mathbf{W}$ and suspected factors in $\mathbf{S}$, further improving the interpretability of the model. It is important to note that while there is further decomposition, the matrices $\mathbf{W}$ and $\mathbf{S}$ can still be interpreted the same way as $\mathbf{W}$ and $\mathbf{H}$ respectively. However, due to the fact that we will now be imposing a constraint on $\mathbf{S}$, $\mathbf{A}$ is needed to perform adjustments on the constrained factors. We propose that $\mathbf{A}$ be initialized as an identity matrix which, through the iterative updating process, will act as a scaling matrix to the rows of $\mathbf{S}$ — a detail which will be shown to be quite significant in Section 4.3.

As previously specified, constraints will now be added to $\mathbf{S}$, 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 $\mathbf{S}$ 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. (2005), an optional constraint regarding the $\mathbf{W}$ matrix has also been added for applications regarding clustering. As previously described in Section 2.4.1, basic NMF resembles a soft K-means clustering algorithm under use of the Frobenius norm as an objective function. Here, the features in $\mathbf{S}$ represent the centroids in the algorithm and the scores in $\mathbf{W}$ are how each observation scores with respect to said centroids. With this in mind, we have allowed the constraint of the first g columns of $\mathbf{W}$ 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 $\mathbf{W}$ and rows of $\mathbf{S}$ 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 the minor dissimilarities are found in the other $q - g$ factors.

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

$$D_F(\mathbf{X} \mid \mathbf{WAS}) = \frac{1}{2} \mid\mid \mathbf{X} - \mathbf{WAS} \mid\mid^2 \tag{3.2}$$

$$\propto \mid\mid \mathbf{X} - \mathbf{WAS} \mid\mid^2 \tag{3.3}$$

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 Section 2.4, it is not feasible to find global minima as the GBR-NMF is not convex in $\mathbf{W}, \mathbf{A}$, and $\mathbf{S}$, which requires an updating process to instead 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 $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{S}$ 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

$$\begin{aligned} D(\mathbf{X} \mid \mathbf{WAS}) &\propto Tr((\mathbf{X}-\mathbf{WAS})(\mathbf{X}-\mathbf{WAS})^T) \\ &= Tr(\mathbf{XX}^T) - 2Tr(\mathbf{WASX}^T) + Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T). \end{aligned} \tag{3.4}$$

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

$$\begin{aligned} \mathcal{L} =& Tr(\mathbf{XX}^T) - 2Tr(\mathbf{WASX}^T) + Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T) \\ &+ Tr(\boldsymbol{\alpha}\mathbf{W}^T) + Tr(\boldsymbol{\beta}\mathbf{A}^T) + Tr(\boldsymbol{\gamma}\mathbf{S}^T). \end{aligned} \tag{3.5}$$

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

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= -2\mathbf{XS}^T\mathbf{A}^T + 2\mathbf{WASS}^T\mathbf{A}^T + \alpha = 0, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{A}} &= -2\mathbf{W}^T\mathbf{XS}^T + 2\mathbf{W}^T\mathbf{WASS}^T + \beta = 0, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{S}} &= -2\mathbf{A}^T\mathbf{W}^T\mathbf{X} + 2\mathbf{A}^T\mathbf{W}^T\mathbf{WAS} + \gamma = 0. \end{aligned} \tag{3.6}$$

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

$$\begin{aligned} (\mathbf{XS}^T\mathbf{A}^T)_{ij}w_{ij} - (\mathbf{WASS}^T\mathbf{A}^T)_{ij}w_{ij} = 0, \\ (\mathbf{W}^T\mathbf{XS}^T)_{ij}a_{ij} - (\mathbf{W}^T\mathbf{WASS}^T)_{ij}a_{ij} = 0, \\ (\mathbf{A}^T\mathbf{W}^T\mathbf{X})_{ij}s_{ij} - (\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij}s_{ij} = 0. \end{aligned} \tag{3.7}$$

Which leads to the following updating rules

$$w_{ij} \leftarrow w_{ij}\frac{(\mathbf{XS}^T\mathbf{A}^T)_{ij}}{(\mathbf{WASS}^T\mathbf{A}^T)_{ij}}, \tag{3.8}$$

$$a_{ij} \leftarrow a_{ij}\frac{(\mathbf{W}^T\mathbf{XS}^T)_{ij}}{(\mathbf{W}^T\mathbf{WASS}^T)_{ij}}, \tag{3.9}$$

$$s_{ij} \leftarrow s_{ij}\frac{(\mathbf{A}^T\mathbf{W}^T\mathbf{X})_{ij}}{(\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij}}. \tag{3.10}$$

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 $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{S}$ are at stationary points (Liu et al., 2011). In order to prove this, we must borrow a lemma stemming from updates to the EM algorithm (Dempster et al., 1977; Wu, 1983) 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 3.2 relevant to w_{ij} and $F'_{w_{ij}}$ is the derivative of such, we then need to prove the following three lemmas.

Lemma 3.2. *The function*

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

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

Lemma 3.3. *The function*

$$G(a, a_{ij}) = F_{a_{ij}}(a_{ij}^t) + F'_{a_{ij}}(a_{ij})(a - a_{ij}^t) + \frac{(\mathbf{W}^T\mathbf{WASS}^T)_{ij}}{a_{ij}^t}(a - a_{ij}^t)^2$$

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

Lemma 3.4. *The function*

$$G(s, s_{ij}) = F_{s_{ij}}(s_{ij}^t) + F'_{s_{ij}}(s_{ij})(s - s_{ij}^t) + \frac{(\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij}}{s_{ij}^t}(s - s_{ij}^t)^2$$

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. (2011), 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 $G(w, w_{ij}^t) \geq F_{w_{ij}}(w)$. This can be done using the Taylor series expansion of $F_{w_{ij}}$:

$$F_{w_{ij}}(w) = F_{w_{ij}}(w_{ij}^t) + F'_{w_{ij}}(w - w_{ij}^t) + \frac{1}{2}F''_{w_{ij}}(w - w_{ij}^t)^2$$

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

$$F_{w_{ij}} = (\frac{\partial D}{\partial W})_{ij} = (-2\mathbf{X}\mathbf{S}^T\mathbf{A}^T + 2\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T)_{ij},$$

$$F''_{w_{ij}} = (\frac{\partial^2 D}{\partial \mathbf{W}^2})_{ij} = 2(\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T)_{jj}.$$

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

$$\frac{(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T)_{ij}}{w_{ij}^t} \geq \frac{1}{2}F''_{ij} = (\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T)_{jj} \tag{3.11}$$

in order to reach Lemma 3.2. Now we have

$$(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T)_{ij} = \sum_{\ell=1}^{q}(\mathbf{W})_{i\ell}(\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T)_{\ell j} \tag{3.12}$$

$$\geq w_{ij}^t(\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A})_{jj} \tag{3.13}$$

which when rearranged proves Equation 3.11. □

3.4 Pseudocode

For the observed $n \times p$ data matrix, $\mathbf{X}$, with k suspected factors in $\mathbf{S}$ 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, $\mathbf{W}$, to be said matrix. Initialize the remaining $n \times (q - c)$ elements sampling from the uniform distribution $U(\min(\mathbf{X}), \max(\mathbf{X}))$. Otherwise, initialize the entire matrix by sampling from the uniform distribution specified above.

2. Initialize $\mathbf{A}$ such that it takes the form of a $q \times q$ identity matrix.
3. Initialize the remaining $(q-k)$ columns of $\mathbf{S}$ sampling from the uniform distribution on the bounds of the known factor values.
4. Calculate $D_F^{(0)}$ using equation 3.2.
5. Set $z = 1$ and enter the updating process:
   (a) Update the unspecified elements of $\mathbf{W}$ using Equation 3.8.
   (b) Update all elements of $\mathbf{A}$ using Equation 3.9.
   (c) Update the unspecified elements of $\mathbf{S}$ using Equation 3.10.
   (d) Calculate $D_F^{(i)}$ using equation 3.2.
   (e) If $D_F^{(z-1)} - D_F^{(z)}$ is less than a specified δ 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 $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{S}$.

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 as lack of progress depends greatly on the value of δ provided. Data sets with large variability may not be able to attain a small epsilon value. Contrarily, if the specified δ 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.

Chapter 4

Applications

4.1 Simulated Data

A simulation was generated in order to test the accuracy 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 $\mathbf{W}$ matrix are set to be the known classification matrix of the model. To add to the simulations, one of the three common features in $\mathbf{S}$ were held as a constrained factor to determine how well it's 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.

$$\mathbf{W} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & w_{1,5} & w_{1,6} & w_{1,7} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & w_{2,5} & w_{2,6} & w_{2,7} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & w_{3,5} & w_{3,6} & w_{3,7} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & w_{4,5} & w_{4,6} & w_{4,7}. \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & w_{5,5} & w_{5,6} & w_{5,7} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & w_{400,5} & w_{400,6} & w_{400,7} \end{pmatrix} \tag{4.1}$$

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{3,3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{4,4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{5,5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{6,6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{7,7} \end{pmatrix} \tag{4.2}$$

$$\mathbf{S} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,2000} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,2000} \\ s_{3,1} & s_{3,2} & \cdots & s_{3,2000} \\ s_{4,1} & s_{4,2} & \cdots & s_{4,2000} \\ \mathbf{s_{5,1}} & \mathbf{s_{5,2}} & \cdots & \mathbf{s_{5,2000}} \\ s_{6,1} & s_{6,2} & \cdots & s_{6,2000} \\ s_{7,1} & s_{7,2} & \cdots & s_{7,2000} \end{pmatrix} \tag{4.3}$$

Entries for $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{S}$ were all created using random uniform distributions and then scaled accordingly such that the column sums of $\mathbf{W}$ are equal to n and the area under the curve of the columns of $\mathbf{S}$ were equal to 1. Thus the diagonal elements of $\mathbf{A}$ were scaled inversely such that the overall reconstructions would remain the same as those prior to scaling.

These data sets were created and decomposed via the GBR-NMF 100 different times in order to ensure accuracy of the results. The residual sum of squares (RSS) comparing the true $\mathbf{S}$ with the estimated $\mathbf{S}$ were then calculated and a mean and standard error of these numbers were taken across all features (found in Table 4.1). This was similarly done with the resulting scores matrix $\mathbf{W}$ and auxiliary matrix $\mathbf{A}$ 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 4.1 displays that the algorithm does exceptionally well in predicting the features in both simulations. Meanwhile, the accuracy of the scores matrix was also quite accurate in comparison to that of the RSS regarding the $\mathbf{A}$ matrix. This is primarily due to the scalings on the two outside matrices which causes the values to become quite close. Therefore, a more fitting comparison would be to the unconstrained model, which is also displayed in Table 4.1. This shows a drastic increase in both the accuracy of the scores and factors in comparison to the original model. 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. (2011) would be inaccurate as a direct comparison is not available. Principal components acquired from PCA are uncorrelated 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. (2011) create different interpretations of the factor matrices.

W constraint held	Standard NMF	
	Average	Standard Error
W RSS	4,786,092,873	19,172,015,537
H RSS	0.000040	0.000010
W constraint held	GBR-NMF	
	Average	Standard Error
WA RSS	2,613,972,969	731,945,188
S RSS	0.000008	0.000004

Table 4.1: Summary of the residual sum of squares (RSS) of both the learned features and scores versus the true features and scores for both standard NMF and GBR-NMF.

4.2 Facial Expression Data

A common research application of nonnegative matrix factorization is data regarding facial expressions (Buciu and Pitas, 2004; Guillamet and Vitria, 2002; Zhi et al., 2010) typically for the identification of certain facial expressions or people. The pain expression data set (Psychological Image Collection at Sterling, 2019) consists of 84 observations of 241×181 pixel facial images. These observations can be further subsetted into the faces of 12 different women; each of which displaying 7 different facial expressions. An average face of each facial expression (aside from neutral) is displayed in row one of Figure 4.1. These six facial expressions are the grouping variable of interest in this data set and are placed as a grouping constraint in the **W** matrix in the GBR-NMF. Along with said 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 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 Figure 4.1 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

Figure 4.1: Row one: Six facial expressions (neutral excluded) averaged over the twelve different women. The facial expressions, in order, are anger, disgust, fear, happiness, pain, and surprise. Row two: Six retrieved unconstrained facial features with respect to grouping matrix corresponding to expressions from the GBR-NMF. Row three: Six retrieved features from the completely unconstrained basic NMF model.

Figure 4.2: An original picture of a woman's face (left) along with reconstructions from the GBR-NMF algorithm (center) and the standard NMF algorithm (right).

row three of Figure 4.1 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 the GBR-NMF in comparison to the basic NMF with respect to the recovered facial expressions. These interpretability benefits could be assumed to result in a decrease in overall model performance. However, as demonstrated in Figure 4.2, 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 are the acquired bases from the model, as well as one's neutral facial expression (ie. the $\mathbf{S}$ matrix). This problem was 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 all of the data regarding a person's facial expressions to be removed from the model and then predicting what their expression would look like; again, using only the neutral facial expression and the bases acquired. 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

Figure 4.3: True facial expression (left) in comparison to GBR-NMF prediction of single missing facial expression (center) and GBR-NMF missing facial expression prediction using only neutral face (right).

shown in Figure 4.3. While it is clear that these reconstructions are not as close to the original image as the results in Figure 4.2, the added benefit of the model is interpretability. In comparison to the conventional NMF, a new facial expression not seen by the algorithm can be predicted due to how easy the factors are to understand. The conventional NMF algorithm would not be able to to this in many cases due because it's factors cannot be interpreted as easily.

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 (Matthews et al., 2015; Harder et al., 2015; Matthews et al., 2010; Paidi et al., 2019). Matthews et al. (2015) 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 (MCF-7), and prostate (LNCaP) tumours. The spectra, overlaid with an average of each cell type, can be seen in the Figure 4.4. 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. (2015) performed a principal component analysis on the spectra and found that the first component closely resembled the metabolite glycogen, a

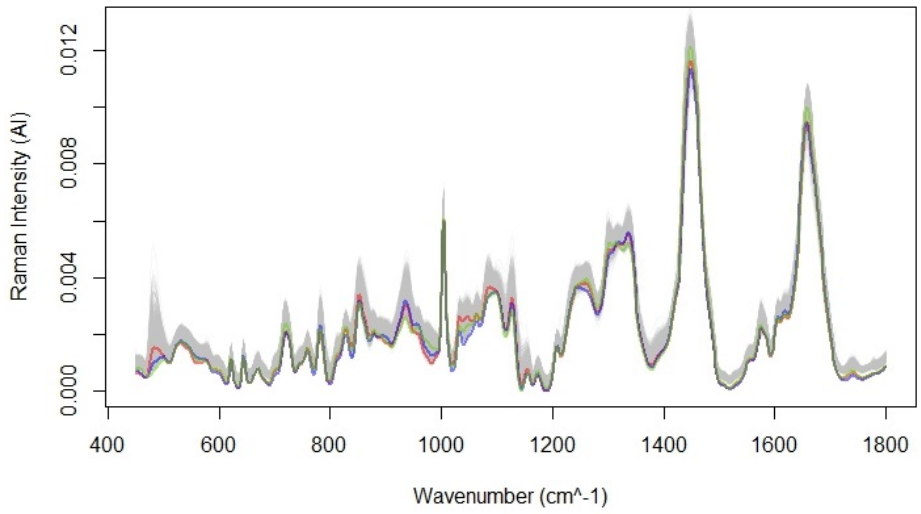

Figure 4.4: Three averaged Raman spectra displaying three different cell lines which include lung (red), breast (blue), prostate (green) tumor cells. Grey lines are the observed Raman spectra. Unit displayed on the y-axis is arbitrary intensity (AI).

key component that is produced in radioresistant cells' response to radiation damage. The aim of this study is to further expand on the work of Matthews et al. (2015) by recovering more biological basis using GBR-NMF. Using the GBR-NMF allows for the constraint of the factor matrix, allowing for the previously recovered glycogen to be specified 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 data set.

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. (2015) 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 matches the physical construction of Raman spectroscopic data. Deng et al. (2020) further illustrates this point by performing NMF on the same

data set and not only discovering glycogen, but also a spectrum very similar to that of a lipid. While using standard NMF is beneficial over PCA, use of the GBR-NMF allows making use of previous results from Matthews et al. (2015) wherein an important chemical, glycogen, was discovered and biologically confirmed. As such, we utilize GBR-NMF by constraining an acquired spectra of pure glycogen within $\mathbf{S}$, and further constrain $\mathbf{W}$ with the known cell lines in the data. We permit one further basis row and group column to be fully unconstrained.

Figure 4.5 provides a scatterplot matrix of the scores from $\mathbf{W}$ 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 tumour 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) fall below a score of 1.5; which matches the result put forward by Matthews et al. (2015) that LNCaP prostate tumours exhibit lower glycogen production in response to radiation than the other two cell lines (H460, MCF7). Furthermore, LNCaP prostate tumours are also known to be more sensitive to radiotherapy than those of H460 and MCF-7. 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 Figure 4.5), 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 MCF-7 cell lines (blue and red). Thus, applying the GBR-NMF method has provided a new spectral component to investigate with regards 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. Further work is required to elucidate the biochemistry of radiation response and is beyond the scope of this thesis.

A natural question that arises is how well the model estimated by GBR-NMF approximates the original spectra. Consider the plot in Figure 4.6 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 an excellent reconstruction.

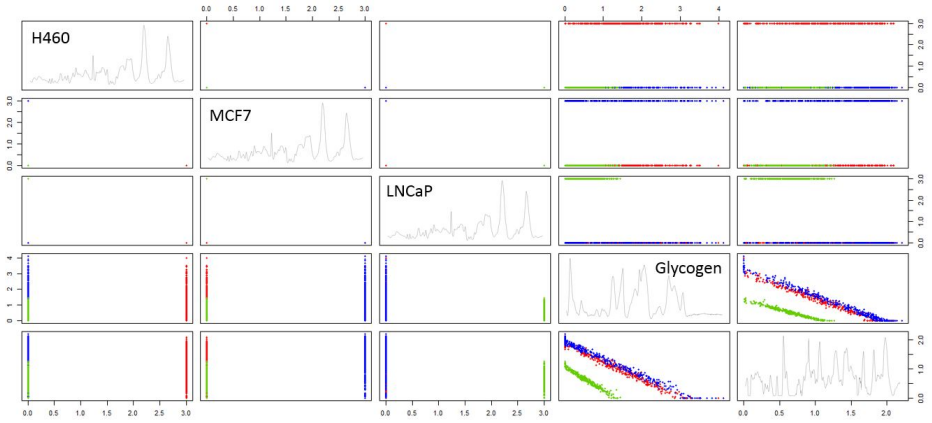

Figure 4.5: Scores on the five (four known/one unknown) chemical bases with respect to the three different cell lines (lung (red), breast (blue), prostate (green)). The Raman spectra in the bottom right corner is that of the unknown feature.

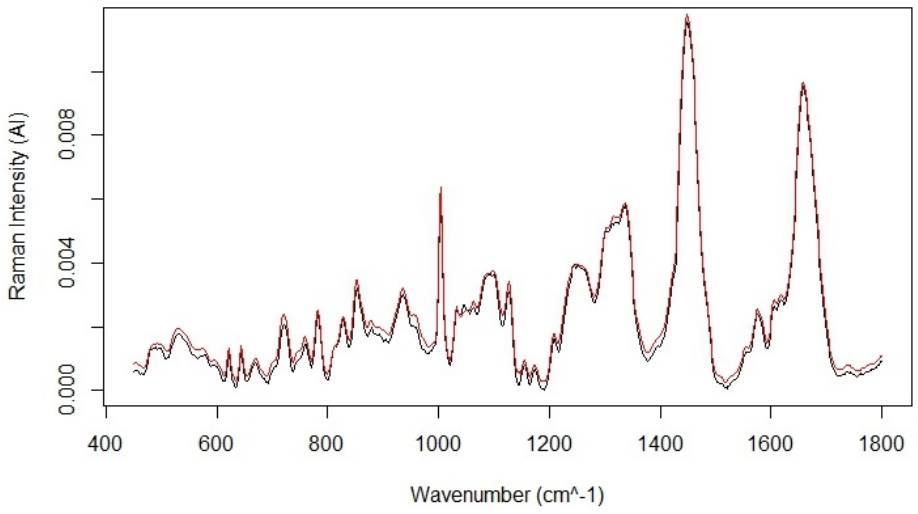

Figure 4.6: An experimentally derived Raman spectrum (black) versus a reconstructed spectrum (red) from the analysis summarized graphically in Figure 4.5.

4.3.1 Raman Spectroscopic Mouse Data

The data set in this example contains 4074 spectra containing 672 different Raman intensities. These spectra were acquired from slices of lungs of mice that had varying degrees of fibrosis and pneumonitis. The degree of sickness in these observations were assessed in accordance with previously established methodologies and were recorded. It is important to note that each observation does not hold its own unique sickness gradings, but instead gradings of the region that the spectrum lies in. These gradings are the proportion of the region that are considered sick and each observation takes these values due to the fact that it is not possible to locate exactly where the spectrum was acquired in the region of interest.

Of the 4074 observations obtained, not all are of the lung tissue described above. This is because some of these spectra taken hit air pockets in the lungs, instead returning the media in which the tissue sample is encased. In order to filter the data set such that only desirable spectra are retained, the GBR-NMF was applied to the data using three different bases - 'good spectra', 'bad spectra', and media. Spectra considered 'good' are those which are representative of Raman spectra if tissue, whereas the media is the solution which the tissues were immersed in during the acquiring of the spectra. The 'bad spectra' are those which resembled the cells to a certain extent, but still displayed peaks which were present in the media. The bases were created by averaging 10 observations considered to take that category; for example, to create the 'good' basis, 10 desirable observations were found in the data and averaged to create said basis. This was also done for both the 'bad' basis, while the media was a pure spectrum of the chemical. Upon initial analysis, a distinct cutoff was determined where any observations with a media score greater than 1 or a 'good' spectrum score less than 1 is removed from the data set. As a result, 2415 of the original 4074 observations were kept. It is important to note that this is not considered outlier removal as not only observations which differ significantly from the rest are being removed. Instead, one should consider it to be group removal, as all spectra which do not fit the desired chemical composition are being removed.

The primary goal with this data was to be able to predict the degree of sickness in a lung tissue sample. More specifically, if one were given a Raman spectrum, could the cell corresponding to that observed spectrum be predicted as sick or not. This can be done through the use of a LASSO logistic regression model, which uses the observation intensities to predict sickness. Based on this model, certain wave numbers are chosen as significant, while others' coefficients are shrunken to zero in order to simplify the model

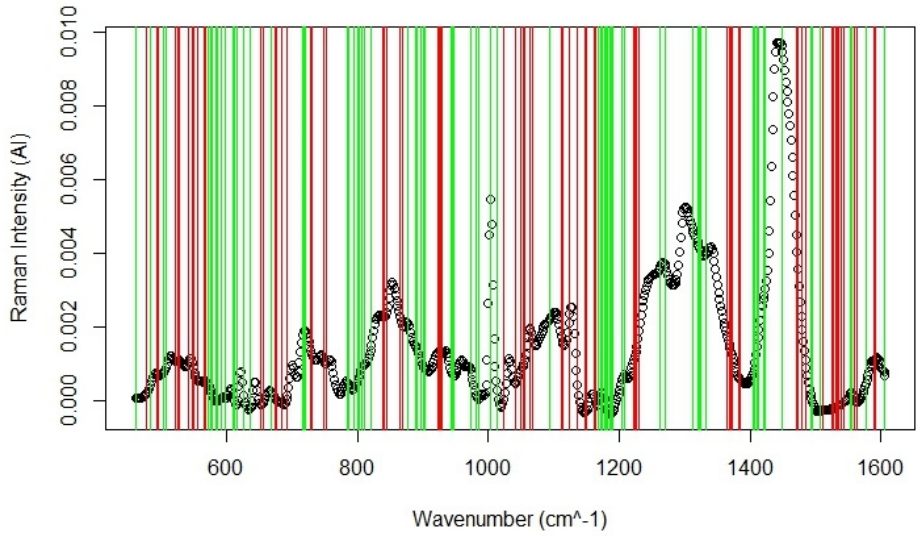

Figure 4.7: Example of statistically significant wavenumbers in predicting fibrotic grade. Statistically significant wavenumbers are shown using both the red and green lines, with red displaying that an increase in intensity results in a decrease in sickness likelihood and green displaying the opposite.

as a whole (see figure 4.7). However, "sickness" is not an actual variable in the data at hand. Each region of observations taken instead has a specific grade between zero and one given to it, which represents the proportion of the observation considered infected. As a result, a threshold needs to be chosen in order to create the response variable of the model. The threshold chosen was 0.5, meaning all observations in a region with a grade below 0.5 is considered healthy and those in regions with grades above are considered sick. This was the initial cutoff point chosen as it was considered the most rational value when choosing a threshold. If more than 50% of the region is considered infected, then all cells are considered disease, causing separation of regions according to the majority of the regions' cell health. While this method is shown to be effective, it is not quite the correct model to be using. Suppose a region has a grading of 0.6. This would then mean that roughly 40% of the observations sampled from this region would not have the disease. Therefore, to classify all observations in this region as infected would be false. Instead, a classification process that has classes other than zero and one should be considered. Herein, multinomial logistic regression is used to classify observations in a multi-class scenario very similarly to the process

Table 4.2: Predicting healthy vs. sick using LASSO logistic regression with a 0.5 grade cutoff.

	Predicted	
True	Healthy	Sick
Healthy:	1654	7
Sick:	4	750

discussed before. This case, however, considers classifying according to the actual grades provided as opposed to sick/healthy. In turn, this then provides a more accurate diagnostic of the observations' level of sickness. It should be noted that this method does not provide an absolute prediction as shown with the previous logistic regression model. For example, suppose an observation is predicted to have a grade of 0.3. Then the accurate interpretation would be that it is predicted that 30% of the cell's region is predicted to be sick. This then could be further interpreted as the cell having 30% probability of being infected. As the grading process can be somewhat subjective, it should also be noted that these grades could be binned into different risk levels. The grades at hand (0,0.05,0.1,0.3,0.5,0.6,0.9,1) can be separated using the two different binning methods outlined below

1. Low (0-0.1), Medium (0.3-0.6), High (0.9-1)
2. Zero (0), Low (0.1-0.3), Medium (0.5-0.6), High (0.9-1)

These methods allow for a more regulated approach among researchers as it does not fall victim to human bias as easily. Once the observations were binned into certain grade groups, multinomial logistic regression was then performed on the data using the spectra as the predictors and the grades as the response. This in theory provides the answer to the question 'can degree of sickness be predicted using Raman spectroscopy?' The resulting classification tables can be found in Tables 4.3 and 4.4.

As shown, the level of sickness can be predicted with a high level of accuracy using the binning processes described above, yielding misclassification rates of 2.15% and 3.15% respectively. This resulted in rates similar to if the original classes were kept. When using the original grades as groups, the multinomial logistic regression model still resulted in a misclassification rate of 2.94%.

This section outlines how GBR-NMF can be effectively used as a method for removing unwanted data. It also displays that with the removal of spectra

Table 4.3: Multinomial logistic regression predictions with respect to binning process number 1.

	Predicted		
True	Low	Medium	High
Low: 0 - 0.1	2050	18	1
Medium: 0.3 - 0.6	21	198	7
High: 0.9 - 1	3	2	115

Table 4.4: Multinomial logistic regression predictions with respect to binning process number 2.

	Predicted			
True	Zero	Low	Medium	High
Zero: 0	1898	15	0	0
Low: 0.1 - 0.3	31	264	1	0
Medium: 0.5 - 0.6	6	18	62	1
High: 0.9 - 1	2	0	2	116

resembling media, level of sickness in mouse lung spectra can be predicted with high levels of accuracy, regardless of how the grades are binned. Multiple binning attempts were made and all of which provided similar results with respect to classification accuracy using multinomial LASSO linear regression. Prediction of a single spectrum's fibrosis grade can be interpreted in two ways; the probability of said cell being sick or the proportion of the section of the lung, which said cell belongs to, that is infected.

Chapter 5

Conclusion

A novel group and basis restricted nonnegative matrix factorization (GBR-NMF) algorithm was developed to input known bases and/or known groups within a nonnegative 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 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 acquired factors from the added group constraint. Advancement over the standard NMF was also proven to be true concerning the motivating data sets as the practicality of the algorithm greatly increase. Being able to specify known factors allows for spectra that may not have been found in the unconstrained form to be discovered.

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 ϵ in Equation 3.1 should also be considered. Namely, the development of new updating procedures in regards to different noise 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 Figure 4.5) from applying GBR-NMF to the data from Matthews et al. (2015) for potential biological importance in indicating tumour 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.

Bibliography

Abdi, H. and L. J. Williams (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics 2* (4), 433–459. 5

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics 44* (1), 197–200. 6

Buciu, I. and I. Pitas (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. 18

Butler, H. J., L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, et al. (2016). Using raman spectroscopy to characterize biological materials. *Nature protocols 11* (4), 664. 2

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39* (1), 1–22. 13

Deng, X., R. Ali-Adeeb, J. L. Andrews, P. Shreeves, J. J. Lum, A. Brolo, and A. Jirasek (2020). Monitor ionizing radiation-induced cellular responses with raman spectroscopy, non-negative matrix factorization, and non-negative least squares. *Applied Spectroscopy*, 0003702820906221. 22

Devarajan, K. (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol 4* (7), e1000029. 8

Ding, C., X. He, and H. D. Simon (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. 9, 11

Feng, X., A. J. Moy, H. T. Nguyen, J. Zhang, M. C. Fox, K. R. Sebastian, J. S. Reichenberg, M. K. Markey, and J. W. Tunnell (2017). Raman active components of skin cancer. *Biomedical optics express 8* (6), 2835–2850. 2

Fonti, V. and E. Belitser (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics 30*, 1–25. 7

Gray, R. M. (1990). Vector quantization. *Readings in speech recognition 1* (2), 75–100. 1

Guillamet, D. and J. Vitria (2002). Classifying faces with nonnegative matrix factorization. In *Proc. 5th Catalan conference for artificial intelligence*, pp. 24–31. 18

Harder, S. J., Q. Matthews, M. Isabelle, A. G. Brolo, J. J. Lum, and A. Jirasek (2015). A raman spectroscopic study of cell response to clinical doses of ionizing radiation. *Applied spectroscopy 69* (2), 193–204. 21

Harman, H. H. (1960). Modern factor analysis. 1

Hartigan, J. A. and M. A. Wong (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28* (1), 100–108. 9

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press. 5

Hilbe, J. M. (2009). *Logistic regression models*. CRC press. 7

Hosmer Jr, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied logistic regression*, Volume 398. John Wiley & Sons. 7

Huang, J., S. Ma, and C.-H. Zhang (2008). The iterated lasso for high-dimensional logistic regression. *The University of Iowa, Department of Statistics and Actuarial Sciences*. 7

Jolliffe, I. (2011). *Principal component analysis*. New York: Springer. 1, 5

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pp. 129–155. Springer. 5

Jolliffe, I. T. (1990). Principal component analysis: a beginner's guide—i. introduction and application. *Weather 45* (10), 375–382. 4

Jolliffe, I. T. and J. Cadima (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374* (2065), 20150202. 4

Kleinbaum, D. G., K. Dietz, M. Gail, M. Klein, and M. Klein (2002). *Logistic regression*. New York: Springer. 7

Kwon, S., S. Han, and S. Lee (2013). A small review and further studies on the lasso. *Journal of the Korean Data & Information Science Society 24*, 1077–1088. 5

Lee, D. D. and H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401* (6755), 788. 1, 7, 8, 10

Lee, D. D. and H. S. Seung (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562. 8

Li, T., C. Ding, and M. I. Jordan (2007). Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 577–582. IEEE. 9

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural computation 19* (10), 2756–2779. 8

Liu, H., Z. Wu, X. Li, D. Cai, and T. S. Huang (2011). Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34* (7), 1299–1311. 9, 10, 13, 14, 17

Matthews, Q., M. Isabelle, S. J. Harder, J. Smazynski, W. Beckham, A. G. Brolo, A. Jirasek, and J. J. Lum (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. 21, 22, 23, 29

Matthews, Q., A. Jirasek, J. Lum, X. Duan, and A. G. Brolo (2010). Variability in raman spectra of single human tumor cells cultured in vitro: correlation with cell cycle and culture confluency. *Applied spectroscopy 64* (8), 871–887. 21

Paidi, S. K., P. M. Diaz, S. Dadgar, S. V. Jenkins, C. M. Quick, R. J. Griffin, R. P. Dings, N. Rajaram, and I. Barman (2019). Label-free raman spectroscopy reveals signatures of radiation resistance in the tumor microenvironment. *Cancer research 79* (8), 2054–2064. 21

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2* (11), 559–572. 4

Pence, I. and A. Mahadevan-Jansen (2016). Clinical instrumentation and applications of raman spectroscopy. *Chemical Society Reviews 45* (7), 1958–1979. 2

Psychological Image Collection at Sterling (2019). 2D face sets - pain expressions data at `pics.stir.ac.uk`. `pics.stir.ac.uk`. Accessed: 2019-09-12. 18

Shevade, S. K. and S. S. Keerthi (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics 19* (17), 2246–2253. 5

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58* (1), 267–288. 5

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine 16* (4), 385–395. 5

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73* (3), 273–282. 5

Wagstaff, K., C. Cardie, S. Rogers, S. Schrödl, et al. (2001). Constrained k-means clustering with background knowledge. In *Icml*, Volume 1, pp. 577–584. 9

Wang, Y.-X. and Y.-J. Zhang (2012). Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering 25* (6), 1336–1353. 1, 8

Wold, S., K. Esbensen, and P. Geladi (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems 2* (1-3), 37–52. 4

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 95–103. 13

Yang, Z., Y. Zhang, Y. Xiang, W. Yan, and S. Xie (2018). Non-negative matrix factorization with dual constraints for image clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 9

Zdunek, R. and A. Cichocki (2007). Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing 87*(8), 1904–1916. 8

Zhai, H. (2018). Variable selection via lasso with high-dimensional proteomic data. *Arts & Sciences Electronic Theses and Dissertations*. 7

Zhi, R., M. Flierl, Q. Ruan, and W. B. Kleijn (2010). Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41*(1), 38–52. 18

Appendix A

Derivation of Updating Algorithm

$$
\begin{aligned}
D_F &= \frac{1}{2} \mid\mid \mathbf{X} - \mathbf{WAS} \mid\mid^2 \\
&\propto \mid\mid \mathbf{X} - \mathbf{WAS} \mid\mid^2 \\
&= (\sqrt{Tr((\mathbf{X} - \mathbf{WAS})(\mathbf{X} - \mathbf{WAS})^T)})^2 \\
&= Tr((\mathbf{X} - \mathbf{WAS})(\mathbf{X} - \mathbf{WAS})^T) \\
&= Tr((\mathbf{X} - \mathbf{WAS})(\mathbf{X}^T - \mathbf{S}^T\mathbf{A}^T\mathbf{W}^T)) \\
&= Tr(\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T - \mathbf{WASX}^T + \mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T) \\
&= Tr(\mathbf{X}\mathbf{X}^T) - Tr(\mathbf{X}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) - Tr(\mathbf{WASX}^T) \\
&\quad + Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T) \\
&= Tr(\mathbf{X}\mathbf{X}^T) - Tr((\mathbf{X}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T)^T) - Tr(\mathbf{WASX}^T) \\
&\quad + Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T) \\
&= Tr(\mathbf{X}\mathbf{X}^T) - Tr(\mathbf{WASX}^T) - Tr(\mathbf{WASX}^T) \\
&\quad + Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T) \\
&= Tr(\mathbf{X}\mathbf{X}^T) - 2Tr(\mathbf{WASX}^T) + Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T)
\end{aligned}
$$

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

$$
\begin{aligned}
\mathcal{L} &= Tr(\mathbf{X}\mathbf{X}^T) - 2Tr(\mathbf{WASX}^T) + Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T) \\
&\quad + Tr(\boldsymbol{\alpha}\mathbf{W}^T) + Tr(\boldsymbol{\beta}\mathbf{A}^T) + Tr(\boldsymbol{\gamma}\mathbf{S}^T)
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= \frac{\partial}{\partial \mathbf{W}}(Tr(\mathbf{X}\mathbf{X}^T) - 2Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) \\
&\quad + Tr(\boldsymbol{\alpha}\mathbf{W}^T) + Tr(\boldsymbol{\beta}\mathbf{A}^T) + Tr(\boldsymbol{\gamma}\mathbf{S}^T)) \\
&= \frac{\partial}{\partial \mathbf{W}}Tr(\mathbf{X}\mathbf{X}^T) - 2\frac{\partial}{\partial \mathbf{W}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + \frac{\partial}{\partial \mathbf{W}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) \\
&\quad + \frac{\partial}{\partial \mathbf{W}}Tr(\boldsymbol{\alpha}\mathbf{W}^T) + \frac{\partial}{\partial \mathbf{W}}Tr(\boldsymbol{\beta}\mathbf{A}^T) + \frac{\partial}{\partial \mathbf{W}}Tr(\boldsymbol{\gamma}\mathbf{S}^T) \\
&= -2\frac{\partial}{\partial \mathbf{W}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + \frac{\partial}{\partial \mathbf{W}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) + \frac{\partial}{\partial \mathbf{W}}Tr(\boldsymbol{\alpha}\mathbf{W}^T) \\
&= -2\frac{\partial}{\partial \mathbf{W}}Tr(\mathbf{X}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) + \frac{\partial}{\partial \mathbf{W}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) + \frac{\partial}{\partial \mathbf{W}}Tr(\boldsymbol{\alpha}\mathbf{W}^T) \\
&= -2\mathbf{X}\mathbf{S}^T\mathbf{A}^T + 2\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T + \alpha \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{A}} &= \frac{\partial}{\partial \mathbf{A}}(Tr(\mathbf{X}\mathbf{X}^T) - 2Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) \\
&\quad + Tr(\boldsymbol{\alpha}\mathbf{W}^T) + Tr(\boldsymbol{\beta}\mathbf{A}^T) + Tr(\boldsymbol{\gamma}\mathbf{S}^T)) \\
&= \frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{X}\mathbf{X}^T) - 2\frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + \frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) \\
&\quad + \frac{\partial}{\partial \mathbf{A}}Tr(\boldsymbol{\alpha}\mathbf{W}^T) + \frac{\partial}{\partial \mathbf{A}}Tr(\boldsymbol{\beta}\mathbf{A}^T) + \frac{\partial}{\partial \mathbf{A}}Tr(\boldsymbol{\gamma}\mathbf{S}^T) \\
&= -2\frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + \frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T)\frac{\partial}{\partial \mathbf{A}}Tr(\boldsymbol{\beta}\mathbf{A}^T) \\
&= -2\frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{W}^T\mathbf{X}\mathbf{S}^T\mathbf{A}^T) + \frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{W}^T\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T)\frac{\partial}{\partial \mathbf{A}}Tr(\boldsymbol{\beta}\mathbf{A}^T) \\
&= -2\mathbf{W}^T\mathbf{X}\mathbf{S}^T + 2\mathbf{W}^T\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T + \beta \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{S}} &= \frac{\partial}{\partial \mathbf{S}}(Tr(\mathbf{X}\mathbf{X}^T) - 2Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) \\
&\quad + Tr(\boldsymbol{\alpha}\mathbf{W}^T) + Tr(\boldsymbol{\beta}\mathbf{A}^T) + Tr(\boldsymbol{\gamma}\mathbf{S}^T)) \\
&= \frac{\partial}{\partial \mathbf{S}}Tr(\mathbf{X}\mathbf{X}^T) - 2\frac{\partial}{\partial \mathbf{S}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{X}^T) + \frac{\partial}{\partial \mathbf{S}}Tr(\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T\mathbf{W}^T) \\
&\quad + \frac{\partial}{\partial \mathbf{S}}Tr(\boldsymbol{\alpha}\mathbf{W}^T) + \frac{\partial}{\partial \mathbf{S}}Tr(\boldsymbol{\beta}\mathbf{A}^T) + \frac{\partial}{\partial \mathbf{S}}Tr(\boldsymbol{\gamma}\mathbf{S}^T)
\end{aligned}
$$

$$
\begin{aligned}
&= -2\frac{\partial}{\partial \mathbf{S}} Tr(\mathbf{WASX}^T) + \frac{\partial}{\partial \mathbf{S}} Tr(\mathbf{WASS}^T\mathbf{A}^T\mathbf{W}^T) + \frac{\partial}{\partial \mathbf{S}} Tr(\boldsymbol{\gamma}\mathbf{S}^T) \\
&= -2\frac{\partial}{\partial \mathbf{S}} Tr(\mathbf{A}^T\mathbf{W}^T\mathbf{XS}^T) + \frac{\partial}{\partial \mathbf{S}} Tr(\mathbf{A}^T\mathbf{W}^T\mathbf{WASS}^T) + \frac{\partial}{\partial \mathbf{S}} Tr(\boldsymbol{\gamma}\mathbf{S}^T) \\
&= -2\mathbf{A}^T\mathbf{W}^T\mathbf{X} + 2\mathbf{A}^T\mathbf{W}^T\mathbf{WAS} + \gamma \\
&= 0
\end{aligned}
$$

Each of these are 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, $\alpha_{ij}w_{ij} = 0$, $\beta_{ij}a_{ij} = 0$, and $\gamma_{ij}s_{ij} = 0$ to the three equations results in the following

$$
\begin{aligned}
(\mathbf{XS}^T\mathbf{A}^T)_{ij}w_{ij} - (\mathbf{WASS}^T\mathbf{A}^T)_{ij}w_{ij} = 0 \\
(\mathbf{W}^T\mathbf{XS}^T)_{ij}a_{ij} - (\mathbf{W}^T\mathbf{WASS}^T)_{ij}a_{ij} = 0 \\
(\mathbf{A}^T\mathbf{W}^T\mathbf{X})_{ij}s_{ij} - (\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij}s_{ij} = 0
\end{aligned}
\tag{A.1}
$$

which leads to the following updating rules

$$
\begin{aligned}
w_{ij} &\leftarrow w_{ij}\frac{(\mathbf{XS}^T\mathbf{A}^T)_{ij}}{(\mathbf{WASS}^T\mathbf{A}^T)_{ij}} \\
a_{ij} &\leftarrow a_{ij}\frac{(\mathbf{W}^T\mathbf{XS}^T)_{ij}}{(\mathbf{W}^T\mathbf{WASS}^T)_{ij}} \\
s_{ij} &\leftarrow s_{ij}\frac{(\mathbf{A}^T\mathbf{W}^T\mathbf{X})_{ij}}{(\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij}}
\end{aligned}
\tag{A.2}
$$

Appendix B

Proof of Algorithm Convergence

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

B.1 Proof of Lemma 3.3

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

$$F_{a_{ij}}(a) = F_{a_{ij}}(a_{ij}^t) + F'_{a_{ij}}(a - a_{ij}^t) + \frac{1}{2}F''_{a_{ij}}(a - a_{ij}^t)^2$$

with $F''_{a_{ij}}$ being the second order derivative of $F_{a_{ij}}$. We can show that

$$F_{a_{ij}} = (\frac{\partial D}{\partial A})_{ij} = (-2\mathbf{W}^T\mathbf{X}\mathbf{S}^T + 2\mathbf{W}^T\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T)_{ij}$$

$$F''_{a_{ij}} = (\frac{\partial^2 D}{\partial \mathbf{A}^2})_{ij} = 2(\mathbf{W}^T\mathbf{W})_{ii}(\mathbf{S}\mathbf{S}^T)_{jj}$$

It is now equivalent to prove

$$\frac{(\mathbf{W}^T\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T)_{ij}}{a_{ij}^t} \geq \frac{1}{2}F''_{ij} = (\mathbf{W}^T\mathbf{W})_{ii}(\mathbf{S}\mathbf{S}^T)_{jj} \tag{B.1}$$

Now we have

$$\begin{aligned}
(\mathbf{W}^T\mathbf{W}\mathbf{A}\mathbf{S}\mathbf{S}^T)_{ij} &= \sum_{\ell=1}^{q}(\mathbf{W}^T\mathbf{W}\mathbf{A})_{i\ell}(\mathbf{S}\mathbf{S}^T)_{\ell j} && \text{(B.2)}\\
&\geq (\mathbf{W}^T\mathbf{W}\mathbf{A})_{ij}(\mathbf{S}\mathbf{S}^T)_{jj} && \text{(B.3)}\\
&\geq \sum_{\ell=1}^{q}(\mathbf{W}^T\mathbf{W})_{i\ell}a_{\ell n}^t(\mathbf{S}\mathbf{S}^T)_{nn} && \text{(B.4)}\\
&\geq a_{ij}^t(\mathbf{W}^T\mathbf{W})_{ii}(\mathbf{S}\mathbf{S}^T)_{jj} && \text{(B.5)}
\end{aligned}$$

Which when rearranged proves Equation B.1. □

B.2 Proof of Lemma 3.4

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

$$F_{s_{ij}}(s) = F_{s_{ij}}(s_{ij}^t) + F'_{s_{ij}}(s - s_{ij}^t) + \frac{1}{2}F''_{s_{ij}}(s - s_{ij}^t)^2$$

with $F''_{s_{ij}}$ being the second order derivative of $F_{s_{ij}}$. We can show that

$$F_{s_{ij}} = (\frac{\partial D}{\partial S})_{ij} = (-2\mathbf{A}^T\mathbf{W}^T\mathbf{X} + 2\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij}$$

$$F''_{s_{ij}} = (\frac{\partial^2 D}{\partial \mathbf{S}^2})_{ij} = 2(\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ii}$$

It is now equivalent to prove

$$\frac{(\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij}}{s_{ij}^t} \geq \frac{1}{2}F''_{ij} = (\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ii} \tag{B.6}$$

Now we have

$$\begin{aligned}((\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ij} &= \sum_{\ell=1}^{q}(\mathbf{A}^T\mathbf{W}^T\mathbf{WA})_{i\ell}(\mathbf{S})_{\ell j} && \text{(B.7)}\\ &\geq (\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ii}s_{ij}^t && \text{(B.8)}\\ &= s_{ij}^t(\mathbf{A}^T\mathbf{W}^T\mathbf{WAS})_{ii} && \text{(B.9)}\end{aligned}$$

which when rearranged proves Equation B.6. □