Non-negative Matrix Factorisation (NMF) introduce by Lee and Seung [1,2], is a model that is used to achieve a dimension reduction on a large complex data matrix to obtain valuable features. It works much like Principal Component Analysis (PCA) but in NMF, each feature in the data matrix, must be greater or equal to zero (non-negative) [3]. Since the introduction, other researchers had successfully implemented NMF in different areas, such as document clustering [4]; information retrieval [5]; facial expression recognition [6]; gene expression analysis [7,8]. The aim is to factorise a non-negative data matrix *A* with dimension *m x n* to produce approximation matrices *WH* with dimensions *m x k* and *k x n* respectively, that is,

(1)

where the k serves as the number of component factors in the model and normally selected so that mn > k(m+n) [1,7].

Regarding this project, Cancer Cell Fraction (CCF) expression values from a set of tumour samples are presented in a matrix *A* with dimension *m x n*, where the rows *m* relates to the band of the CCF, the columns *n* relates to the tumour samples and entries are the total values of the CCF that falls within the bands. After applying NMF on the CCF expression value matrix *A* to *WH*, where *W* has dimension *m x k*, such that *k* columns determine the number of topics in the expression and where H has dimension *k x n,* such that *n* columns reveal the weight of the expression of the tumour sample. Furthermore, the *W* can be described as a feature matrix and *H* as the coefficient matrix.

**Implementation of the Model**

To implement the NMF model, a data matrix was created using the 500 sample files from the Subclonal Structure after phase one. Each sample file has a corresponding CCF values. Thus, all the CCF values from each file were counted and arranged in bands that start from 0.1 to 1 with step 0.1. The shape of the data matrix is10 x 500 (row x column). The figure below shows the average (mean) count of the total CCF value from each of the tumour samples.

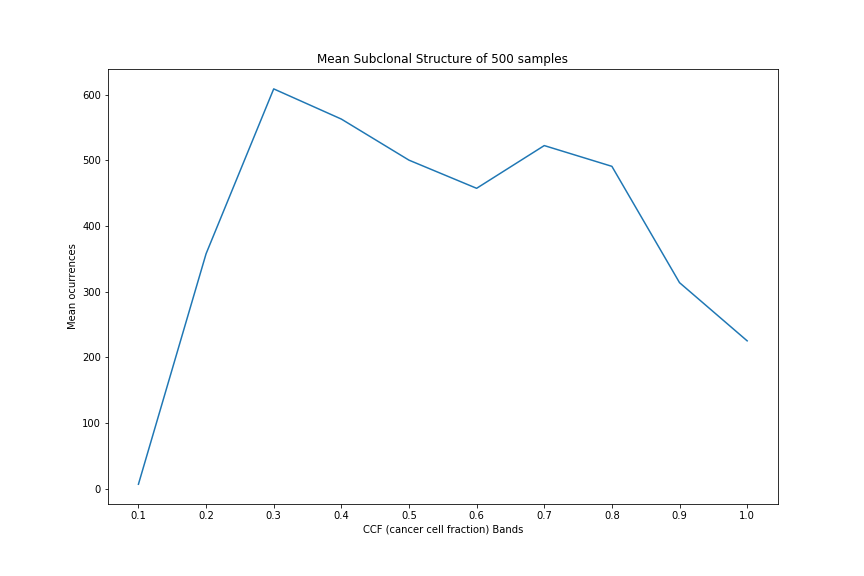


Figure X, *The CCF against the mean number of ocurrences from the the subclonal structure results during the first phase.*

For NMF, the data matrix can be defined as *A* with dimension *m x n*, where the row *m* defines the 10 different bands *{0.1, 0.2, ...., 1}* and where column *n* defines the value of tumour mutation from the 500 sample.

Regarding this project, data pre-processing such as using Count Vectorise to return the count of the data is not needed as the data matrix is already in numbers. To factorise the data matrix *A* to get the product of *WH*, an objective function must be defined to measure the approximation and the reconstruction error. This function can be designed by adjusting the distance between *A* and the multiplication result *WH* [7]. The commonly used method for measuring the distance is the squared *Frobenius norm* a branch of the *Euclidean norm* [8]. The *Frobenius norm* [1,9],

(2)

A different method for measuring the distance is *Kullback-Liebler* [1,7]. The commonly used process for adjusting the *W* and *H* in Equation (2) to reduce the objection function is to use the multiplicative update rules to iterate between the *W* and *H* until convergence [1]. The rules are as follow,

(3)

,

In order to get a better factorisation of the data matrix *A* into a feature matrix *W* and coefficient matrix *H*, we had to perform a few analyses such as tuning the parameters to find the less reconstruction error and estimating the number of components (*k*) for the model to determine the optimal parameters for the NMF model. The first phase of the analysis was to estimate the number of components to use for the NMF model. This was achieved by measuring the explained variance from the data matrix *A* to reveal how many variances (information) can be associated with each of the components. We measure this more directly by creating a method to utilise a metric library from Scikit-learn to get the explain variance score from the NMF model and the data matrix *A*.

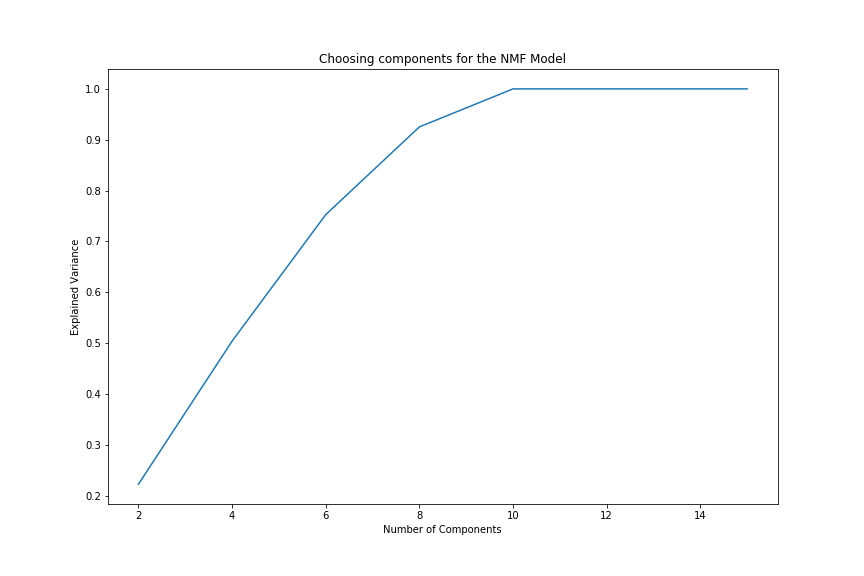


Figure 1, estimating the number of components (k) for NMF model through explain variance.

The judging from the figure above, it evidently shows that 6 to 9 components are needed to approximately explain 92% of the variance. However, 99.95% to be specific of the variance can be explained using components from 10. The next phase of the analysis was to tune the model to determine the parameters with least reconstruction error. To achieve this, we created a method that takes in the data matrix *A* and some NMF model parameters and returned the reconstruction error. Considering this project the NMF parameters are, n\_components, this states the number of components to use; init, this defines the initialisation method utilised for the model; max\_iter, this determines the maximum iterations before timeout; random\_state, this is stated to check reproducibility; l1\_ratio, this defines the L1 and L2 penalties to regularise the model; alpha, this is used to control the strength of the regularization [10]. Subsequently, a form a grid search was performed to find the optimal parameters for the model. The table below shows the three smallest and largest reconstruction errors for different parameter combinations.

Table 1, shows three combinations with the smallest and largest error.

|  |  |
| --- | --- |
| Three Smallest Error | |
| Parameters | Reconstruction Error |
| (10, ‘nndsvd’, 50, 0, 0, 0) | 1.024160 |
| (10, ‘nndsvd’, 50, 1, 0.5, 0.75) | 1.235355 |
| (10, ‘nndsvd’, 50, 1, 1, 0.75) | 1.653756 |
|  | |
| Three Largest Error | |
| (4, ‘nndsvda’, 2, 1, 1, 0) | 61975.362915 |
| (4, ‘nndsvda’, 2, 0, 1, 0.25) | 61967.598853 |
| (2, ‘nndsvda’, 2, 1, 1, 0) | 61965.296342 |

Prior searching for the optimal parameters for the model, it was explicit that most of the variance can be explained using 10 components. As a result, the maximum number of components used for tuning the model was 10. Furthermore, when tuning the model, it clearly shows that the higher the component the lower the reconstruction error of the model will be (Figure ).

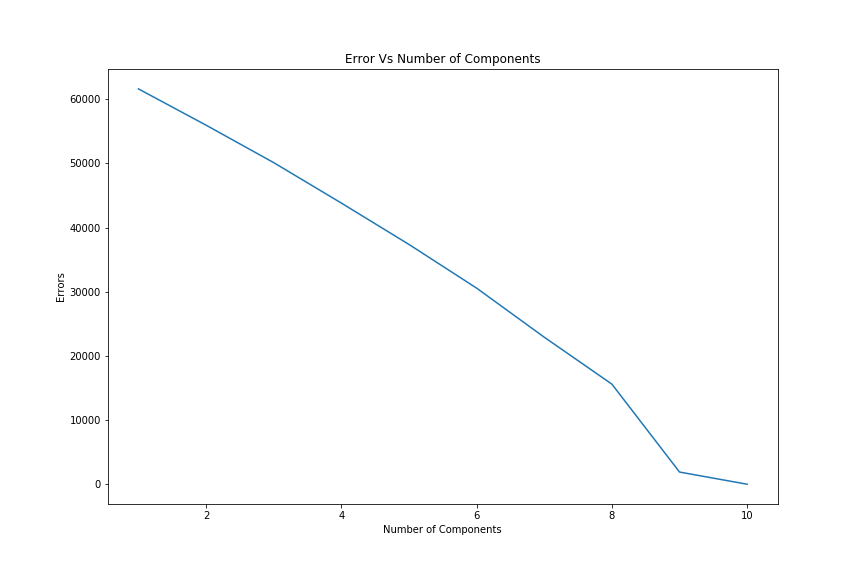


Figure 2, the number of components against the reconstruction error.

The NMF model was applied to the data matrix A utilising the optimal parameters identified during the analysis phase. The number of components was set to 10; init = ‘nndsvd’, which means the initialization method utilised for the model is "Nonnegative Double Singular Value Decomposition"; max\_iter = 50; random\_state = 0; l1\_ratio = 0, this means the L2 (Frobenius Norm) penalty for regularisation; alpha = 0. After factorising the data matrix A with dimension 10 x 500 (m x n), the model produced approximation matrices WH, where W is the feature (topic) matrix with dimension 10 x 10 (m x k) and H the coefficient matrix with dimension 10 x 500 (k x n). For the feature matrix, the rows are the bands of the CCF and the columns are the number of components (topics) used for the model. While for the coefficient matrix, the rows are the number of components and columns are the tumour samples.

the feature matrix was further normalised using a library part of the Scikit-learn preprocessing libraries to get a unit value.

Figure 3-4 are plots of the feature matrix for five documents and coefficient matrix respectively.

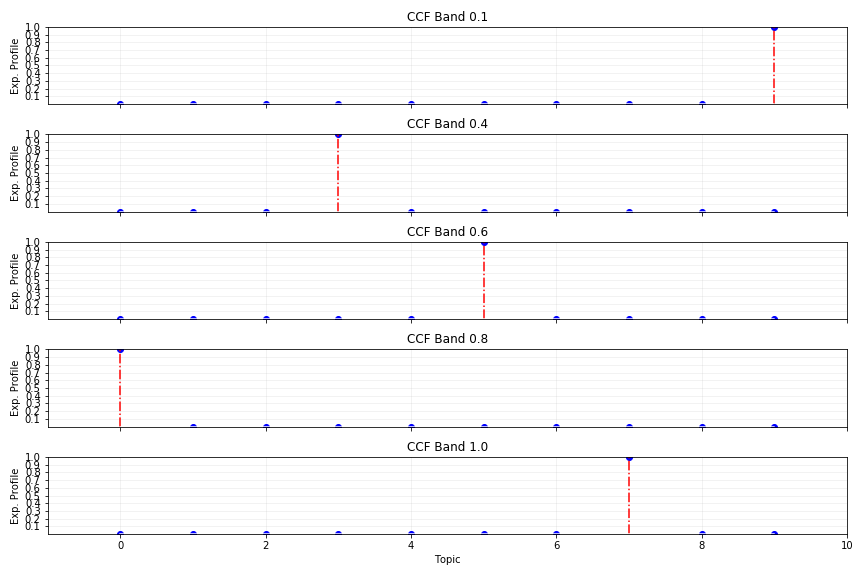


Figure 3, five CCF (Cancer Cell Fraction) bands for the feature (Topic) matrix

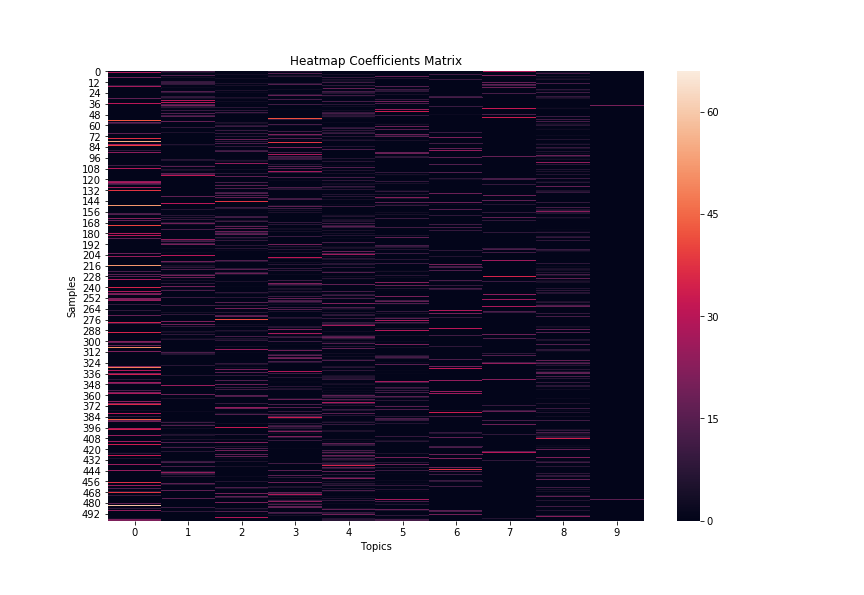


Figure 4, plot for the coefficient matrix

**Reference**

1. Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788-791.

2. Lee D, Seung H. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems. 2001; 556-562.

3. Müller A, Guido S. Introduction to Machine Learning with Python. 1st ed. Sebastopol: O’Reilly Media, Inc.; 2016.

4. Shahnaz F, Berry M, Pauca V, Plemmons R. Document clustering using nonnegative matrix factorization. Information Processing & Management. 2006;42(2):373-386.

5. Tsuge S, Shishibori M, Kuroiwa S, Kita K. Information Retrieval Using Non-negative Matrix Factorization. IEEJ Transactions on Electronics, Information and Systems. 2004;124(7):1500-1506.

6. Salim W, Santika D. Human Facial Expression Recognition Using Two-dimensional Non-negative Matrix Factorization. Procedia Engineering. 2012;50:758-767.

7. Devarajan K. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. PLoS Computational Biology. 2008;4(7):e1000029.

8. Lee C, Mudaliar M, Haggart D, Wolf C, Miele G, Vass J et al. Simultaneous Non-Negative Matrix Factorization for Multiple Large Scale Gene Expression Datasets in Toxicology. PLoS ONE. 2012;7(12):e48238.

9. Decomposing signals in components (matrix factorization problems) — scikit-learn 0.20.1 documentation [Internet]. Scikit-learn.org. [cited 3 December 2018]. Available from: <https://scikit-learn.org/stable/modules/decomposition.html#nmf>

10. Sklearn.decomposition.NMF — scikit-learn 0.20.2 documentation [Internet]. Scikit-learn.org. [cited 3 December 2018]. Available from: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>

**Not used yet**

Maisog J. Non-negative Matrix Factorization: Assessing Methods for Evaluating the Number of Components, and the Effect of Normalization Thereon [Masters]. Georgetown University; 2009.