Scope
Volume 15 Number 01 March 2025

Topic Modeling: A Review via Nonnegative Matrix Factorization

[2*]Francis Attah; [1]Hong Keat Yap; [2]Wah June Leong; [3]Sie Long Kek
[1]Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, 43200 Kajang, Selangor Malaysia
[2]Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor Malaysia
[3]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia Pagoh Campus, Jln Panchor, 84600 Panchor, Johor Malaysia

Corresponding Author: **Francis Attah**

Abstract: This work offers a thorough analysis of topic modeling using Nonnegative Matrix Factorization (NMF), a potent method that is frequently employed to identify significant themes and patterns in sizable unstructured datasets. NMF has become a vital tool in domains such as document clustering, social network analysis, bioinformatics, and natural language processing by breaking down data into interpretable components. Important NMF variations are covered in this review, such as Nonnegative Matrix Tri-Factorization (NMTF), factorization objective techniques, constrained NMF, and algorithmic improvements. We go over how NMF can be used in a variety of contexts, evaluate its effectiveness, and take into account how constraints can improve clustering quality.

Keywords: Topic modeling, nonnegative matrix factorization, social network analysis, document clustering.

1. Introduction

Over the last few decades, topic modeling has emerged as a crucial technique for classifying and comprehending enormous volumes of unstructured text data. Social media databases and digital libraries now house enormous volumes of information due to the increasing digitization of literature across various disciplines. The researchers need advanced automated tools that can analyze and extract important themes, similar to how people interpret information, in order to make sense of this data. Several algorithms are used in the topic modeling technique, which assists in identifying and classifying the thematic structure in sizable document collections. In order to provide a meaningful synopsis of the main subjects covered in the documents, topic modeling aims to uncover latent thematic structures within a text corpus (Blei, 2012). Microblog stream is one of its many analytical applications (Huang et al, 2017), bioinformatics (Liu et al, 2016), filtering documents (Li et al, 2018), social data (Hong and Davison, 2010) and a lot more. Researchers have examined topic modeling from a variety of perspectives over the past 20 years, drawing from a broad range of fields such as computer science, biology, statistics, mathematics, and neuroscience. Numerous studies and publications have resulted from this, concentrating on different topic modeling techniques like Nonnegative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and others.

In summary, this paper makes the following contributions: it designs the NMF and its structural classification hierarchy methods. Presentations of the NMF objective function and computational algorithms were made. Additionally, the review will provide insightful information about Nonnegative Matrix Factorization research, specifically in the fields of co-clustering and word clustering. The rest of the paper is organized as follows: Section 2 discusses NMF and its structural classification; Section 3 introduces different measure functions and the corresponding NMF computational algorithm; and Section 4 addresses issues with extra constraints. Section 5 presents the quantitative evaluation of NMF, and Section 6 provides the review's conclusion.

2. NMF and Structural Classification

Nonnegative Matrix Factorization (NMF) is a popular method for dimensionality reduction and data decomposition. It is a subset of linear algebra algorithms that are intended to reveal the underlying or hidden structures within data. With all entries nonnegative, it uses the product of two lower-rank matrices, M and N, to approximate a given nonnegative matrix X. The original data is compactly represented by this decomposition, where M stands for the bas is components and N for the coefficients required to reconstruct the data using these components. According to the research work by (Lee and Seung, 1999)., NMF is especially appealing because it can create sparse, parts-based representations of the data, improving interpretability. In accordance with (Burred's 2014) research, start with a $(u \times v)$ nonnegative data matrix X, where u and v represent the number of rows and columns respectively, and r is the reduced rank, NMF looks for a $(u \times r)$ matrix M and a $(r \times v)$ matrix N such that.

$$X \approx MN.$$

Reconstruction error, which is frequently quantified in terms of the Frobenius norm or Kullback-Leibler divergence, was to be minimized by this factorization problem:

$$\min_{M,N} \|X - MN\|^2 \quad \text{Subject to} \quad M \geq 0, N \geq 0.$$

NMF is particularly helpful for applications where only additive combinations have meaning because of its non negativity constraint. To extract latent topics, for example, NMF is applied to document-term matrices in text mining. Each column of M represents a topic, and the matrix N shows the degree to which each document is associated with these topics (Xu et al, 2003). Likewise, in image processing, NMF can break down images into basic characteristics, like discrete facial elements in face recognition software (Guillamet et al, 2002). Because NMF is nonnegative, these decompositions are easier to understand because they represent specific portions of the original data rather than arbitrary combinations.

Many data analysis applications have adopted it due to its simplicity and clarity (Chen et al, 2017).

NMF is a potent method for condensing the enormous datasets produced by contemporary web and computer technologies into a lower-dimensional space. This method, which is customized for the particular context of the data, is used to efficiently find underlying themes, salient characteristics, or latent variables. In this paper, we describe our approach to categorizing **NMF techniques according to various standards, such as:** The fundamental NMF model which serves as the foundation for all other NMF variations. There are nine subclasses of the **Constrained NMF**: Standard NMF, in which both factorized matrices are required to contain nonnegative elements. With sparse NMF, one or both factorized matrices are subject to sparsity constraints. The factor matrices are subject to orthogonality constraints thanks to orthogonal NMF. Smooth NMF creates smoother factors by applying smoothness constraints. In order to prevent over fitting, regularized NMF uses regularization terms. Semi-NMF permits one of the factor matrices to have both positive and negative values by easing the non negativity constraint on it. In order to enforce relationships between data points, graph regularized NMF introduces a graph-based regularization. Manifold NMF integrates various learning methods into NMF, such as graph Laplacian regularization. By using a kernel trick, Kernel NMF makes it possible to capture non-linear relationships.

Objective-based Factorization is divided into three main categories: Frobenius norm-based NMF, Kullback-Leibler divergence, and Generalized Divergence NMF. In contrast, the algorithm-based update can be categorized into Multiplicative Update algorithms, Projected Gradient Descent, Alternating Least Squares (ALS), Probabilistic NMF, and Hierarchical NMF. The systematic NMF is classified into five sections: Nonnegative Tensor Factorization (NTF), Non-negative Matrix Set Factorization (NMSF), Convolutive NMF, Weighted NMF, and Nonnegative Matrix Tri-Factorization (NMTF), which decomposes a data matrix into three factor matrices.

3. Measure Functions of NMF

There are various challenges when using NMF in real-world applications, especially because of its NP-hardness (Vavasis, 2009). As a result, standard nonlinear optimization techniques are used in many algorithms, which frequently result in convergence at stationary points as opposed to global optima. Iteratively minimizing a cost function that evaluates the similarity between X and the product MN by updating the factor matrices alternately is a more workable approach. The first way to frame the problem is to introduce a scalar error measure, D(X, MN), which is also referred to as the objective function, the cost, or the loss function. This illustrates how the factorized product MN and the input matrix X differ from one another. Either a divergence or a distance metric may be used for this measurement. The Frobenius norm, the (generalized) Kullback-Leibler divergence (KL) divergence, and the Itakura-Saito (IS) divergence are the three most often used error functions in literature. These are shown in Eq (1), (2), and (3), respectively (Burred, 2014).

$$D_{EU}(X, MN) = ||X - MN||^2, \quad (1)$$

$$D_{KL}(X, MN) = \sum_u \sum_v \left(X_{uv} \log \frac{X_{uv}}{(MN)_{uv}} - X_{uv} + (MN)_{uv} \right) \quad (2)$$

$$D_{IS}(X, MN) = \sum_u \sum_v \left(\frac{X_{uv}}{(MN)_{uv}} - \log \frac{X_{uv}}{(MN)_{uv}} - 1 \right) \quad (3)$$

The total squared error between the elements of the original data matrix and its approximation is measured by the Frobenius norm-based NMF. This method of NMF works well for problems where the goal is to minimize reconstruction error and is perfect for situations where maintaining the overall scale and structure of the data is crucial. It does not favor high or low values; instead, it handles all errors equally. On the other hand, KL divergence is an objective function that quantifies how different two probability distributions are from one another. In the context of NMF, it contrasts the product matrices MN with the original data matrix X, which is thought to represent a distribution of some kind. The type of data and the particular objectives of the factorization task frequently influence the choice between the various objective functions. As an illustration, (Yan et al, 2013) NMF based on KL divergence performed poorly. This could be because the sparse data caused it to over fit the data. In contrast, (Vangara et al, 2021) suggested a Semantic NMF based on KL divergence in order to tackle the problem of determining the appropriate number of topics, also known as the latent dimension of a corpus. The technique accurately determines the number of topics in text corpora and enables them to determine the latent dimension of the semantically enhanced topics based on their stability. In order to ascertain the unknown number of latency, a random procedure based on the initial TF-IDF and SPPMI matrices and custom clustering were used. In the literature, there are additional types of divergence measures of dissimilarity-based NMF, such as Itakura-Saito (IS) divergence (Fevotte et al, 2009). Bregman divergences are the class to which it belongs (Banerjee et al, 2005) as well as a limit case of the β-divergence (Cichocki et al, 2006), the α-divergence (Hoseinipour et al, 2023), among many others. When decomposing piano spectrograms, IS-NMF works especially well at precisely separating out elements associated with low residual noise and string hammer strikes. The Euclidean or KL divergence methods, on the other hand, frequently ignore or drastically distort these subtle features (Fevotte et al, 2009).

According to Lee and Seung (2001), the iterative process of optimizing the factor matrices to minimize the discrepancy between the original matrix and its approximation is known as the computational algorithmic update procedure with NMF. The NMF algorithm iteratively modifies the values in matrices M and N in order to minimize an objective function, most often the Kullback-Leibler divergence or the Frobenius norm. The goal function is convex in either matrix M or N, but not both. So, repeated multiplicative updates (MU) are used to optimize the system. The process of applying the Frobenius norm to two factor matrices entails holding the other matrix constant while minimizing the first matrix under constraints, as shown in equations (4) and (5) below. Once convergence is achieved, either a limited number of iterations or another stopping condition is met, or the factor matrices provide a good approximation of the original data, this process is repeated by switching the roles of the matrices.

$$D_{EU}(X, MN) = \min_{M>0} ||X - MN||^2 \quad (4)$$

$$D_{EU}(X, MN) = \min_{N>0} ||X - MN||^2 \quad (5)$$

This MU approach is favored because it is straightforward and easy to use, but depending on initialization. It may converge slowly and become trapped in a local minima (Gills, 2014), (Lee and Seung, 2001), (Sim et al., 2018, 2019). It is important to note that some NMF solutions are not unique. It could be because of numerical instability. Prior research has proposed an efficient method to deal with this: normalizing each column vector of matrix M (Peng et al, 2020b), (Deng et al, 2023), (Peng et al., 2020a) to establish a unit norm. Normalization is done using either L_1 or L_2 norms. Additionally, to reduce the possibility of division by zero, a small positive constant is typically included in the denominator of a complete MU algorithm for the standard NMF (Farial, 2004), (Mirzal, 2014). Similar to MU, the Alternating Least Square (ALS) (Liu et al, 2013) algorithmcan iteratively approximated NMF by breaking it down into the product of two lower-rank nonnegative matrices using the method. In order to solve each factor matrix for a least-squares minimization, ALS fixes one factor matrix at a time while updating the other. It uses various objective functions for optimization, including Kullback-Leibler divergence, the Frobenius norm, and other kinds of divergence. It frequently converges more quickly than multiplicative updates. Gradient Descent-based NMF Algorithms (Peng et al., 2020b) iteratively minimize the objective function by updating the factor matrices following its negative gradient. This general optimization method is used with NMF.

4. NMF Problems with Additional Constraints

To improve a conventional NMF problem's performance or customize it for a particular application, additional constraints are frequently required. Without giving the factor matrices any extra structure, conventional NMF merely reduces the reconstruction error. Nonetheless, more constraints can (a) enhance interpretability (Pei et al, 2019) (b) In fields such as topic modeling or image processing, constraints such as sparsity, which promotes zero values in factors, aid in the creation of parts-based representations that are simpler to understand. (c) Data Structure Reflection (Shaodi et al, 2019), (Meng et al, 2019); by applying structure-preserving constraints, like graph regularization or orthogonality, it is possible to better capture the inherent structure of some datasets, such as smoothness, clustering tendencies, or graph relationships. (d) Increase Robustness (Sha and Diao, 2022); in real-world situations, NMF may converge to unwanted solutions due to noisy or incomplete data. Regularization ($L_{2,1}$) and orthogonality are two examples of constraints that can reduce overfitting or improve the stability of the decomposition. (e) Promote the use of discriminative factors (Peng et al. 2017) is a restriction that can improve the differentiation of different data classes, which is helpful for supervised tasks like classification. For example, discriminative constraints on factor matrices can enhance NMF's capacity to identify class differences in labeled data.

5. Quantitative Assessment of NMTF

Our choice for the quantitative analysis of NMF methods is nonnegative matrix tri-factorization (NMTF) techniques, specifically looking for word clusters in the works of (Lee and Seung, 1999), (Ding et al, 2006) as well as (Kong et al., 2011). Traditional NMF was expanded to handle more complex structures, especially for co-clustering tasks, with the development of NMTF. Early research on tri-factorization concepts was impacted by the findings of (Ding et al, 2006), who applied NMTF for clustering and data co-clustering tasks in the mid-2000s. The method was first made popular in fields such as collaborative filtering, bioinformatics, and text mining, where three-factor decomposition aids in the discovery of latent structures in data matrices with row and column clusters. The NMTF extends the fundamental NMF to the product of three factor matrices, M, G, and N, given a nonnegative data matrix X where

$$X \approx MGN.$$

The algorithm is as discussed in section 3.

For the quantitative analysis, we selected three data sets which come from Kaggle (Classic 3, BBC News) and Git Hub (Medical): (i) Classic3 dataset, see (Hoseinipour et al, 2023), a popular dataset in the fields of machine learning, text mining, and information retrieval, especially for assessing algorithms for topic modeling, document clustering, and classification. It contains documents from three traditional text collections from the SMART Information Retrieval System: articles and abstracts from the medical field, documents pertaining to space and aviation research, and information science topics. The dataset is comprised of 3 classes, 4303 dimensions, and 3891 sample documents. (ii) Textual data from medical records, scientific publications, and other documents pertaining to healthcare are included in the medical dataset. In healthcare settings, this dataset is frequently used to investigate clustering or classification; possible topics include patient history, symptoms, procedures, or diagnoses. The dataset has three classes and 4499 sample documents, among other notable features. (iii) Comprising text documents from the BBC, the BBC News dataset, see (Gao et al, 2015) is usually categorized into five areas: tech, politics, entertainment, business, and sports. Its well-defined topics make it popular in text mining and natural language processing research, where it is helpful for assessing clustering algorithms. The total number of documents is 1490.

Pre-processing is the process of turning unstructured text into a form that can be analyzed. The text was first standardized by changing it to lowercase, eliminating punctuation and numbers, and eliminating common stop words that didn't add much context. Next, the text is tokenized, which divides it into discrete words or tokens. By grouping related terms together and condensing words to their root forms, stemming or lemmatization is used to further reduce redundancy. Then, using models like Bag-of-Words, which counts word frequencies, or TF-IDF, which allocates weights based on term distinctiveness across documents, a vocabulary of unique terms is generated across all documents and numericized. Additionally, normalization is used to scale the data consistently in order to get it ready for tasks like topic modeling and clustering. When evaluating NMTF methods, one usually looks at the quality of the clusters that are produced and how well the factorization resembles the original data matrix. primarily entails evaluating the robustness, interpretability, and clustering accuracy.

Table 1: Top Word clusters (WC) from four algorithms using a medical dataset

NMFK=3					
WC1:	seizure	aggressive	ana	pleura	genitalia
WC2:	jaw	Sub periosteal	cup	Vicod in	multivitamin
WC3:	woman	sperm	voiding	flap	showed
NMF-$L_{2,1}$K=3					
WC1:	history	patient	mg	pain	normal
WC2:	patient	placed	procedure	incision	right
WC3:	artery	coronary	left	catheter	right
NMTFK=3					
WC1:	congenital	gallbladder	correlation	lost	hydralazine
WC2:	noncontrast	bit	insert	disorder	extensive
WC3:	cabg	advanced	dull	acute	lopressor
ONMTFK=3					
WC1:	placed	procedure	patient	incision	using
WC2:	artery	coronary	left	right	catheter
WC3:	history	patient	mg	pain	past

Examining the top word clusters (WCs) in each of the four algorithms, as shown in Table 1, shows that each algorithm aims to represent the datasets—in this case, the medical dataset—in a clear, human-like manner. When closely examined, the top words in each of the word clusters (WCs) produced by ONMTF and NMF-$L_{2,1}$ create distinct word clusters that each brilliantly capture the distinctive features of their respective clusters. The topics of patient incision procedures are highlighted in ONMTF by WC1, coronary artery problems are the focus of WC2, and patient pain history is covered in WC3. Even though the word clusters from other algorithms are fairly clear, they don't have the clear division that ONMTF and NMF-$L_{2,1}$ do. The additional constraints incorporated into these two approaches might be the cause of this improved distinction.

In summary, all of the algorithms can be used for clustering in natural language processing tasks like image processing, text mining, social network analysis, and bioinformatics. However, more significant clusters can be produced by imposing particular restrictions according to the goals of the study. The quality and interpretability of the results are improved by these constraints, which also help make semantic themes more clear, particularly when working with large datasets.

6. Conclusions

NMF has gained popularity recently as a clustering and co-clustering technique in natural language processing and machine learning. By classifying large volumes of unstructured data into comprehensible topics or themes, it is especially useful for organizing and labeling such information. This study examines a number of structured classification techniques, such as constrained NMF, objective-based factorization, algorithmic variations, and NMTF, while providing a thorough analysis of topic modeling using NMF. In addition, a comprehensive assessment of topic modeling with NMTF is given, along with specific information about its uses in domains like image processing, text mining, document clustering, social network analysis, and bioinformatics.

Acknowledgement

The corresponding author is funded by the Tertiary Education Trust Fund (TETFUND) of Nigeria, under reference number TETF/UNIV/NASARAWA/TSAS/2022.

References

1. Banerjee, A., Dhillon, I.S., and Ghosh, J. (2005).Clustering with Bregman divergences. Journal of Machine Learning Research, 6:1705–1749.
2. Blei, D. M. (2012). Probabilistic topic models. Communication of the ACM, 55(4):77–84.
3. Burred, J. (2014).Detailed derivation of multiplicative update rule for NMF. Computer Science, Mathematics, ID: 49323490.
4. Chen, B.W., Ji, W., Rho, S., and Gu, Y. (2017). Supervised collaborative filtering based on ridge alternating least squares and iterative projection pursuit. IEEE Access, 5:6600–6607.
5. Cichocki, A., Zdunek, R., and Amari, S. (2006). Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06), pages 32–39.
6. Deng, F., Zhao, Y.,Pei,J., Wang, S., and Yang, X. (2023). Nonnegative matrix factorization constrained by multiple labelled spanning trees for label propagation. Information Sciences, 648. 119616.
7. Ding, C., Li, T., Peng, W., and Park, H. (2006).Orthogonal non-negative matrix tri-factorization for clustering. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 126 - 135.
8. Fevotte,C., Bertin, N., and Durrieu, J.-L.(2009).Non negative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Computation, 21:793-830.
9. Gao, H., Cai, W., Nie, F., and Huang, H. (2015). Robust capped norm nonnegative matrix factorization. ACM.
10. Gills, N. (2014). The why and how of nonnegative matrix factorization. Machine Learning, 2:257–291.

11. Guillamet, D., Bernt, Schiele, and Vitrià, J. (2002). Analyzing non-negative matrix factorization for image classification. International Conference on Pattern Recognition, pages 116–119.
12. Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. Proceedings of the First Workshop on Social Media Analytics, SOMA, 10:80–88.
13. Hoseinipour, S., Aminghafari, M., and Mohammadpour, A. (2023). Orthogonal parametric non-negative matrix tri-factorization with α-divergence for co-clustering. Expert System with Application, 231(4).
14. Huang, J., Peng, M., Hua, W., Cao, J., Gao, W., and Zhang, X. (2017). A probabilistic method for emerging topic tracking in microblog stream. World Wide Web, 20(2):325–350.
15. Kong, D., Ding, C., and Huang, H. (2011).Robust nonnegative matrix factorization using $l_{2,1}$ norm. pages 673–682.
16. Lee, D. D. and Seung, H.S. (1999). Learning the parts of objects by nonnegative matrix factorization. Nature, 401(6755):788–791.
17. Lee, D. D. and Seung, H. S. (2001). Algorithm for nonnegative matrix factorization. Advances in Neural Information Processing System, 13:556–562.
18. Li, C., Chen, S., Xing, J., Sun, A., and Ma, Z. (2018).Seed-guided topic model for document filtering and classification. ACM Transactions on Information Systems, 37(1):9.
19. Liu, H., Li, X., and Zheng, X. (2013).Solving non-negative matrix factorization by alternating least squares with a modified strategy. Data Min Knowl Disc, 26:435-451.
20. Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. Springer Plus, 5:1608.
21. Meng, Y., Shang, R., Shang, F., Jiao, L., Yang, S., and Stolkin, R. (2019).Semi-supervised graph regularized deep NMF with bi-orthogonal constraints for data representation. IEEE Transactions on Neural Networks and Learning Systems, 31(9):1–14.
22. Mirzal, A. (2014). A convergent algorithm for orthogonal nonnegative matrix factorization. Journal of Computational and Applied Mathematics, 260:149–166.
23. Pei, Y., Fletcher, G., and Pechenizkiy, M. (2019).Joint role and community detection in networks via $L_{2,1}$ norm regularized nonnegative matrix tri-factorization.pages 168–175.
24. Peng, L., Peng, J. and Fan, J. (2017).$L_{2,1}$ norm and Hessian regularized nonnegative matrix factorization with discriminability for data representation. Applied Sciences, 7(10):1013.
25. Peng, S., Ser, W., Chen, B., Sun, L., and Lin, Z. (2020b). Robust nonnegative matrix factorization with local coordinate constraint for image clustering. Engineering Applications of Artificial Intelligence, 88:103354.
26. Sha, X. and Diao, N. (2022). $L_{2,1}$-norm-based robust feature extraction method for fault detection. ACS Omega, 7(48): 43440–43449.
27. Shaodi, G., Li, H., and Luo, L. (2019). Constrained dual graph regularized orthogonal nonnegative matrix tri-factorization for co-clustering. Mathematical Problems in

Engineering. ID. 7565640.

28. Sim, H. S., Leong, W. J., Chen, C. Y., & Ibrahim, S. N. I. (2018). Multi-step spectral gradient methods with modified weak secant relation for large scale unconstrained optimization. Numerical Algebra, Control and Optimization, 8(3):377-387.
29. Sim, H. S., Leong, W. J., & Chen, C. Y. (2019). Gradient method with multiple damping for large-scale unconstrained optimization. Optimization Letters, 13(1):617-632.
30. Vangara, R., Bhattarai, M., Skau, E., Chennupati, G., Djidjev, H., Tierneys, T., Smith, J.P., Stanev, V.G., and Alexandrov, B.S. (2021). Finding the number of latent topics with semantic non-negative matrix factorization. IEEE Access, 9:117217–117231.
31. Vavasis, S. (2009).On the complexity of nonnegative matrix factorization. SIAM J. Optimization, 20(3):1364–1377.
32. Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pages 267–273.
33. Yan, X., Guo, J., Liu, S., Cheng, X., and Wang, Y. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. Proceedings of the 2013 SIAM International Conference on Data Mining, pages 749–757.