SDS
Statistica e
Data Science

Palermo, 11-12 April 2024

Proceedings of the Statistics and Data Science 2024 Conference

New perspectives on Statistics and Data Science

Edited by

Antonella Plaia – Leonardo Egidi
Antonino Abbruzzo

Proceedings of the SDS 2024 Conference
Palermo, 11-12 April 2024
Edited by: Antonella Plaia – Leonardo Egidi - Antonino Abbruzzo
-
Palermo: Università degli Studi di Palermo.

ISBN Ebook (Pdf)
978-88-5509-645-4

Contents

Preface

The development of large-scale data analysis and statistical learning methods for data science is gaining more and more interest, not only among statisticians, but also among computer scientists, mathematicians, computational physicists, economists, and, in general, all experts in different fields of knowledge who are interested in extracting insight from data. Cross-fertilization between the different scientific communities is becoming crucial for progressing and developing new methods and tools in data science. In this respect, the Statistics & Data Science group of the Italian Statistical Society has organized its 3rd international conference held in Palermo on the 11st and 12nd of April 2024, attended by over 100 researchers from different scientific fields. A collection of the presented papers is available in the present Proceedings showing a huge variety of approaches, methods, and data-driven problems, always tackled according to a rigorous and robust scientific paradigm.

The Statistics & Data Science group

Palermo, April 11st and 12th, 2023

Antonella Plaia - Leonardo Egidi - Antonino Abbruzzo
Editors

Determining the optimal number of clusters through Symmetric Non-Negative Matrix Factorization

Stavolo Agostino[1], Grassia Maria Gabriella, Marino Marina, Mazza Rocco, Paesano Simone, Sacco Dario

Abstract Cluster analysis, as a form of unsupervised learning, has been developed to group observations by leveraging application-specific similarity measures. This study investigates matrix factorization techniques, with a specific focus on analyzing lexical tables within the framework of term-document matrices. Symmetric Non-Negative Matrix Factorization (SNMF) takes center stage as an effective tool for clustering operations. The primary challenge addressed is the automated determination of the optimal number of clusters.

Key words: clustering, symmetric non-negative matrix factorization

1 Introduction

The transition from analog to digital data, driven by advancements in information technology and the Internet's growth, has generated vast data volumes across domains. Textual datasets pose challenges due to high dimensionality and sparsity, demanding significant computational resources and impacting result generalizability. Analyzing large corpora faces high dimensionality issues, as lexical tables form large, sparse matrices. This data noise must be minimized to reduce computational complexity and ensure reliable results in associative text relationship

[1] Stavolo Agostino, University of Naples Federico II, agostino.stavolo@unina.it
Grassia Maria Gabriella, University of Naples Federico II, mariagabriella.grassia@unina.it
Marino Marina, University of Naples Federico II, marina.marino@unina.it
Mazza Rocco, Univeristy of Bari Aldo Moro, rocco.mazza@uniba.it

analyses. Various approaches exist for matrix dimensionality reduction and automatic information extraction from document collections.

Factorial-based Approach: Utilizes factorization techniques like LCA [9] and LSA [3] based on GSVD to minimize the number of terms needed for describing documents in a low-rank vector space.

Network-based Approach: Represents matrices as graphs to visualize contextual relationships between terms [12], overcoming Bag-of-Words encoding limitations. Community detection methods are commonly used [5].

Probabilistic Approach: Addresses dimensionality reduction through probabilistic models such as topic modeling (pLSA [6] and LDA [1]), enabling the discovery of underlying topics within a corpus.

Non-Negative Factorization (NMF): Utilizes NMF to achieve reduced dimensionality through additive combinations of original terms, with applications in various fields [10]. However, its effectiveness in clustering is limited for nonlinear cluster structures, leading to the introduction of symmetric Non-Negative Matrix Factorization (SNMF).

The contribution suggests utilizing consensus clustering to determine the ideal number of semantic clusters for delineating the primary themes within Ursula von der Leyen's communication, employing Symmetric Non-Negative Matrix Factorization.

2 Symmetric Non-Negative Matrix Factorization

The objective of Symmetric Non-Negative Matrix Factorization (SNMF) is to approximate a non-negative symmetric matrix A by expressing it as the product of a non-negative matrix H and its transpose H^T, resulting in $A \approx HH^T$.

During clustering operations, outcomes are obtained directly by marking the number of columns corresponding to the maximum value in each row of H [15], making H also known as the clustering assignment matrix. SNMF, falling under the category of soft graph clustering, relaxes the orthogonality constraint on the H matrix, resulting in approximately orthogonal matrices and fuzzy clustering results, distinguishing it from traditional hard graph clustering methods. Due to its advantages in data clustering, SNMF has gained significant attention, leading to the development of various algorithms. Researchers have recognized SNMF's superior performance in clustering tasks [14] [2], and it has been noted for its effectiveness in extracting topics from lexical matrices [18]. Recent studies [16], have shown that SNMF outperforms both k-means and spectral clustering methods, indicating its potential in diverse applications. SNMF has found success not only in clustering tasks but also in community detection within graphs, as demonstrated [11]. The formalization of SNNMF in the following equation:

$$\min_{H \geq 0} ||A \approx HH^T||_F^2 \qquad (1)$$

It is considered a generalized form for solving clustering objectives [7]. Involving the minimization of the Frobenius norm, the non-negative matrix H plays a crucial role, with dimensions n × k, where k is the number of required clusters. It is essential for capturing the cluster structure. SNMF is essentially a form of clustering on graphs, demonstrating similarities to kernel k-means and spectral clustering, given its relaxation of non-negativity orthogonality on H. The performance of SNMF clustering relies on the construction of the similarity matrix A. Various methods, proposed by [8], include considering the affinity matrix Ã, where elements represent weights of edges between data points, and constructing A based on ratio correlation or normalized cut. The SNMF algorithm is designed for decomposing completely positive similarity matrices, presenting challenges when negative eigenvalues are involved. To address this, the single-loss weight SNMF model was developed by [4].

3 Automatic number of clusters in SNMF

A significant challenge for clustering methods is determining the number of clusters in the data. For the SNMF we use the consensus clustering to define the automatic number of *k* clusters [13]. Consensus clustering formalizes the concept of amalgamating diverse clusters into a unified representative, or consensus, to highlight shared structures across various datasets and unveil substantial differences among them. The objective of consensus clustering is to identify a representative consensus that reflects the provided cluster of a particular dataset.

Formally, the vector ID_q comprises labels, $ID_{qj} \in [1, ..., k]$, indicating the cluster memberships for each row in X. The binary connectivity matrix, B_q, is derived from the values in the ID_q vector, where B_{ij} equals 1 if the i*th* and j*th* elements belong to the same cluster and 0 otherwise. Formally, $B^{(q)}{}_{(i,\ j)}$ = {1 if x_i and x_j belong to the same cluster, 0 otherwise. The consensus matrix $M^{(k)}$ is computed for a fixed number of clusters *k* by averaging all binary connectivity matrices:

$$M^{(k)} = \frac{\sum_{q=1}^{r} B^q}{r} \tag{2}$$

Each element in $M^{(k)}$ (known as the consensus index), indicates how frequently the i*th* and j*th* elements are clustered together and *r* is the number of resamples. A perfect clustering results in a consensus matrix with only 0 or 1 values. Deviations from these values quantify the stability of clusters for a given number of clusters, and by comparing consensus matrices $M^{(k)}$ for various $k \in [k_l, k_u]$, the number of clusters corresponding to the minimum deviation from a perfect consensus is identified as the potential number of clusters in the actual data.

4 Empirical study

To empirically showcase the practical value of consensus clustering in automatically identifying the optimal number of clusters in the SNMF, we conducted an analysis on Instagram posts made by Ursula von der Leyen, the President of the European Commission. Our objective was to delineate the primary topics of discussion spanning from her inauguration year in 2018 to January 1, 2024. We use Crowd Tangle to extract 708 posts by the official profile of the European President. The preprocessing phases of cleaning textual data involve several essential steps to prepare the text for analysis. Firstly, normalization ensures consistency by converting text to a standard format. This may involve converting all letters to lowercase and removing punctuation marks. Next, tokenization breaks the text into individual words or tokens, separating them into distinct units. Removing stopwords is another crucial step, where common words that don't carry significant meaning, such as "the" or "and" are eliminated to reduce noise. Lemmatization helps further refine the text by reducing words to their root form. Stemming removes affixes from words, while lemmatization maps words to their base form [12]. After that we create the vocabulary, and we apply the cosine similarity to create a term-term matrix. In Figure 1 we use the consensus clustering to cosine similarity matrix and identify five topics (Table 1), according to the lower value of the consensus measure.

Figure 1: Optimal number of clusters

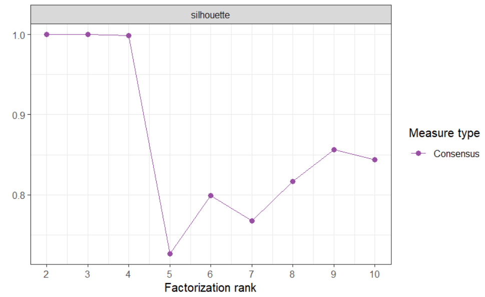

Table 1: Semantic clusters of Ursula von der Leyen communication

Clusters	***Words***
Balkan policies	Albania, Machedonia, Heads_state, implement, government, earthquake, solidarity, balkan, call, trip
Ukrainian war	Respond, dependency, consumer, russian, belarus, propose, attack, sanction, diplomatic, Ukraine
Green economy	Green, plan, invest, digital, resilience, hydrogen, recovery, project, next_generation, sustainable
Refugees	Foster, family, housing, child, lose, house, damage, home, responsibility, refugee,
Covid-19	Vaccination, vaccine, pandemic, Europe, order, dose, announce, test, safe_vaccine

Considering the top ten terms appearing in the cluster, the first topic delves into Balkan support policies, specifically focusing on the comparison between Macedonia and Albania, alongside the management of earthquake funds. The discussion likely centers around the initiatives and strategies implemented by international organizations and governments to provide support and assistance to the Balkan region, with a specific focus on Macedonia and Albania. The second topic involves the ongoing conflict between Ukraine and Russia and the role of Europe in maintaining stability and balance in international relations. The conflict has been a significant geopolitical issue, with implications for regional security and global politics. Europe, as a key player in international affairs, has a vested interest in resolving the conflict peacefully and preventing further escalation. Europe's role may include diplomatic efforts, economic sanctions, peacekeeping initiatives, and support for Ukraine's sovereignty and territorial integrity. The third theme pertains to government policies and initiatives aimed at promoting a green economy and transitioning to renewable energy sources. A green economy focuses on sustainable development, environmental protection, and reducing carbon emissions to combat climate change. Policies related to this may include incentives for renewable energy production, investments in clean technologies, regulations to reduce greenhouse gas emissions, and support for sustainable practices in various sectors such as transportation, agriculture, and industry. The use of renewable energies, such as solar, wind, hydroelectric, and geothermal power, is a key component of transitioning to a green economy, as it reduces reliance on fossil fuels and mitigates environmental impacts. The last topic defines the vaccination strategy of Covid-19 pandemic. The focus is on vaccination efforts in Europe, encompassing discussions on vaccine procurement, distribution, and the scientific campaign to support vaccination efforts.

5 Conclusions and limitations

The paper delves into various matrix factorization techniques, with a specific focus on utilizing lexical tables to unveil latent themes within textual data. Its primary objective is to effectively decompose these matrices, with a particular emphasis on symmetric non-negative matrix factorization (SNMF), renowned for its effectiveness in clustering tasks. The central aim of the study is to develop an automated method for determining the optimal number of clusters within a lexical matrix. Among the methodologies investigated, consensus clustering emerges as a promising approach, backed by the analysis results, supported by the results of the analysis, which showed the main themes of Ursula von der Leyen's communication on Instagram. However, the paper recognizes the computational challenges associated with consensus clustering, especially when handling large datasets or a high number of resamples, as it can lead to a substantial computational burden.

References

1. Blei D.M., Ng A.Y., Jordan M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022 (2003)
2. Chen, B. W.: Symmetric nonnegative matrix factorization based on box-constrained half-quadratic optimization. IEEE Access, 8, 170976-170990. (2020)
3. Deerwester, S. C., Dumais S. T, Landauer T. K., Furnas G. W., Harshman. R. A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6):391–407. (1990).
4. Ding, C., He, X., Simon, H. D.: On the equivalence of nonnegative matrix factorization and spectral clustering. In Proceedings of the 2005 SIAM international conference on data mining (pp. 606-610). Society for Industrial and Applied Mathematics. (2005).
5. Fortunato, S.: Community detection in graphs. Phys. Rep. 486, 75–174 (2010)
6. Hofmann T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1–2):177–196. (2001)
7. Kuang D, Ding C, Park H.: Symmetric nonnegative matrix factorization for graph clustering. In: Proceedings of SIAM international conference on data mining (SDM), pp 106–117. (2012)
8. Kuang D., Choo J., Park H.: Nonnegative matrix factorization for interactive topic modeling and document clustering. In Partitional Clustering Algorithms (pp. 215-243). Springer, Cham. (2015)
9. Lebart L., Salem A., Berry L.: Correspondence Analysis of Lexical Tables., Exploring Textual Data. Text, Speech, and Language Technology, vol 4. Springer, Dordrecht. (1998)
10. Lee D. D., Seung H. S.: Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791. (1999).
11. Luo, X., Liu, Z., Jin, L., Zhou, Y., Zhou, M.: Symmetric nonnegative matrix factorization-based community detection models and their convergence analysis. IEEE Transactions on Neural Networks and Learning Systems, 33(3), 1203-1215. (2021).
12. Misuraca, M., Scepi, G., Spano, M.: Network-Based Dimensionality Reduction for Textual Datasets. In Models for Data Analysis: SIS 2018, Palermo, Italy, June 20–22 (pp. 175-190). Cham: Springer International Publishing. (2018).
13. Vangara, R., Rasmussen, K. Ø., Chennupati, G., Alexandrov, B.: Determination of the number of clusters by symmetric non-negative matrix factorization. In Big Data III: Learning, Analytics, and Applications (Vol. 11730, pp. 104-113). SPIE. (2021).
14. Wu, W., Jia, Y., Kwong, S., Hou, J.: Pairwise constraint propagation-induced symmetric nonnegative matrix factorization. IEEE transactions on neural networks and learning systems, 29(12), 6348-6361. (2018)
15. Yan, W., Zhang, B., Yang, Z., Xie, S.: Similarity learning-induced symmetric nonnegative matrix factorization for image clustering. IEEE Access, 7, 166380-166389. (2019).
16. Yan, X., Guo, J., Liu, S., Cheng, X., Wang, Y.: Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In proceedings of the 2013 SIAM International Conference on Data Mining (pp. 749-757). Society for Industrial and Applied Mathematics. (2013).