MATH 5473 Project 2

LI Haobo
hliem@connect.ust.hk

SHENG Rui
rshengac@connect.ust.hk

CHEN Zixin
zchendf@connect.ust.hk

TENG Fei
fteng@connect.ust.hk

Abstract

Understanding the evolution and trends in a specific research field is crucial for advancing knowledge. However, with the ever-increasing volume of academic papers, researchers face significant challenges in efficiently navigating and identifying the prevailing topics. To address this issue, we present a novel approach aimed at automatically extracting and organizing topics from the extensive NIPS paper dataset. By leveraging our methodology, researchers can gain valuable insights into emerging themes, influential ideas, and crucial knowledge gaps. This not only facilitates a deeper understanding of the field but also fosters the progress of research by enabling researchers to make informed decisions and contribute to the cutting-edge developments.

1 Introduction

The hotspot of academic writing and paper acceptance is a major concern for researchers. There has been a noticeable shift in research focus over the past few decades, making it essential for researchers to stay up-to-date with current trends and recent topics Bishop and Nasrabadi [2006]. By doing so, researchers can ensure that their work is relevant and appealing to the contemporary academic community, increasing their chances of success in the publishing process. However, the sheer volume of research papers across various areas makes it impractical to analyze the progression of topic shifts. To address this issue, we propose a visual analytics approach aimed at helping researchers explore the evolution of research topics over the past ten years in accepted papers within a certain area using the NIPS dataset.

Specifically, we first count the appearance of words for each year in the dataset and generate a word frequency matrix. Since the number of different words that appeared in NIPS is huge within the period from 1987 to 2015, and the presence of a large number of infrequent words causes the matrix to be sparse, we applied a dimension reduction technique to project the original word frequency matrix into a 2-dimensional or 3-dimensional space. The reduced space can be visualized by a spatial graph directly, allowing us to intuitively understand the similarities and differences between different texts. Moreover, we can apply clustering algorithms to this dimension reduced space to divide the papers into several groups. Each group, containing a set of papers, can represent a topic about the machine learning area. As a result, we apply the Silhouette Coefficient to evaluate the similarity between the clusters computed by the clustering algorithm and describe each topic by summarizing the frequent words within the set of papers. Additionally, we conduct an experiment to compare the effectiveness of 8 dimension reduction techniques by examining the clustering results.

Overall, our contributions to this work can be summarized as follows:

1) We propose a visual analytics system for mining the hot research topics in the machine learning area for NIPS-accepted papers.

2) We compare the performance of eight dimension reduction techniques on the hotspot topic mining task. The results show that UMAP is the most promising technique for the task.

The interactive visualization can be found in: https://cinderd.github.io/MATH5473$_Final_Project/$. And the link for the source code is: https://github.com/CinderD/MATH5473$_Final_Project/tree/main$.

The contribution of our group members are:

- Li Haobo: Result analysis, Report writing: Experiments, Slides making, Presentation
- Chen Zixin: Coding: Visualization, Silhouette analysis, Presentation
- Sheng Rui: Coding: Dimension reduction, Report writing: Discussion, Presentation
- Teng Fei: Report writing: Intro and Method, Slides making, Presentation

2 Methods

2.1 Feature Extraction

Different topics have different focus words. Therefore, based on the NIPS dataset from 1987 to 2015, we extracted features by using the frequency of word occurrences as topic features. Since the papers contained in the dataset were all accepted by the NIPS conference and from the machine learning area, there exists seldom outliers for word frequencies. However, since the papers are collected in a relatively large time interval (from 1987 to 2015) and 11,463 words are counted, the word frequency matrix is sparse and requires preprocessing before topic extraction.

2.2 Dimension Reduction

Since the word frequency matrix records 11,463 unique words, a considerable portion of the vocabulary did not appear in many years, resulting in the sparsity of the matrix. It is difficult to use for downstream tasks. Therefore, we used dimension reduction techniques to address the problem. To find out which dimensionality reduction method is the most appropriate, we compared eight methods, such as Principal Component Analysis (PCA), to project the word frequency features into a lower dimensional space. Besides, we used visualization to display the dimension reduced data in a two-dimensional or three-dimensional space, providing an intuitive understanding of the similarities and differences between different texts. Additionally, reducing the number of sparse features allows clustering algorithms to effectively use features and compute clusters more accurately.

2.3 Topic extraction

Since the NIPS dataset does not provide topic labels, unsupervised learning clustering algorithms are suitable for extracting and summarizing hot topics. As the NIPS dataset contains machine learning papers, which have few outlier samples, and the features are in vector form, we used the K-means algorithm Lloyd [1957], which is easy to converge, as the clustering algorithm. After clustering, we used the Silhouette Coefficient to calculate the similarity of each cluster and extracted the top 20 most frequent words in each cluster to describe it.

3 Dimension Reduction Methods

For dimension reduction, we employed eight methods as follows. We compared their performance in downstream tasks.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) Hotelling [1933] is a widely used method for reducing the dimensionality of high-dimensional data. It does this by creating new, uncorrelated variables, called principal components, that capture the most important patterns in the original data. PCA is a linear method that maximizes variance, making it effective for data reduction and visualization.

3.2 t-distributed Stochastic Neighbor Embedding

t-distributed Stochastic Neighbor Embedding (t-SNE) Van der Maaten and Hinton [2008] is a non-linear dimensionality reduction technique that is particularly effective for visualizing high-dimensional datasets. It works by converting similarities between data points into probabilities, and then tries to minimize the difference between these probabilities in the high-dimensional and low-dimensional spaces.

3.3 Gaussian Random Projection

Gaussian Random Projection Bingham and Mannila [2001] is a method that uses random projections to reduce the dimensionality of data. It works by projecting the original data onto a lower-dimensional subspace using a random matrix with elements drawn from a Gaussian distribution. This technique is particularly effective for high-dimensional datasets with a large number of features.

3.4 Kernel PCA

Kernel PCA Schölkopf et al. [2005] is a non-linear extension of PCA that can capture non-linear patterns in high-dimensional data. It works by mapping the original data into a higher-dimensional feature space using a kernel function, and then performing PCA in this space. Kernel PCA is particularly effective for datasets with non-linear relationships between the features.

3.5 Fast Independent Component Analysis

Fast Independent Component Analysis (Fast ICA) Hyvärinen [2013] is a method for blind source separation, which aims to extract independent sources from a set of mixed observations. It works by maximizing the non-Gaussianity of the source signals, which can be useful for identifying hidden patterns in high-dimensional data.

3.6 Sparse PCA

Sparse PCA Zou et al. [2006] is a variant of PCA that encourages sparsity in the principal components. It works by adding a penalty term to the PCA objective function that penalizes non-zero coefficients in the principal components. This can be useful for identifying important features in high-dimensional datasets.

3.7 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) Lee and Seung [2000] is a method for decomposing a high-dimensional data matrix into two non-negative matrices, where the columns of one matrix represent the basis vectors and the rows of the other matrix represent the coefficients of these basis vectors. NMF is particularly effective for datasets where the features are non-negative and can be used for feature extraction and clustering.

3.8 Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) McInnes et al. [2018] is a non-linear dimensionality reduction technique that is similar to t-SNE. It works by constructing a low-dimensional manifold that preserves the topological structure of the high-dimensional data. UMAP is particularly effective for datasets with complex non-linear relationships between the features.

4 Experiments

4.1 Dimension Reduction

We conducted an experiment to compare the effectiveness of eight different dimension reduction methods in reducing the dimensionality of sparse word data and removing redundant information such as highly correlated or noisy features. In addition to improving computational efficiency, dimension

	2	3	4	5	6	7	8	9	10
UMAP	0.28038	0.31398	0.34937	0.33965	0.37173	0.38427	0.3828	0.37585	0.3796
NMF	0.34011	0.32037	0.35316	0.32854	0.29451	0.30228	0.28285	0.29533	0.29169
S-PCA	0.2887	0.33466	0.34789	0.2728	0.27141	0.26443	0.25834	0.26574	0.26998
F-ICA	0.24986	0.30885	0.33374	0.31522	0.28035	0.2695	0.27046	0.26615	0.26857
K-PCA	0.33769	0.33697	0.34276	0.34971	0.33397	0.3257	0.30687	0.31112	0.3063
GRP	0.22104	0.21584	0.21968	0.216	0.20764	0.20637	0.20771	0.20361	0.20974
TSNE	0.25212	0.24675	0.25236	0.25468	0.26337	0.26572	0.25293	0.24948	0.24557
PCA	0.27732	0.33537	0.34634	0.27386	0.29446	0.26532	0.25749	0.26476	0.26661

Figure 1: Silhouette Scores for different methods and numbers of clusters(k)

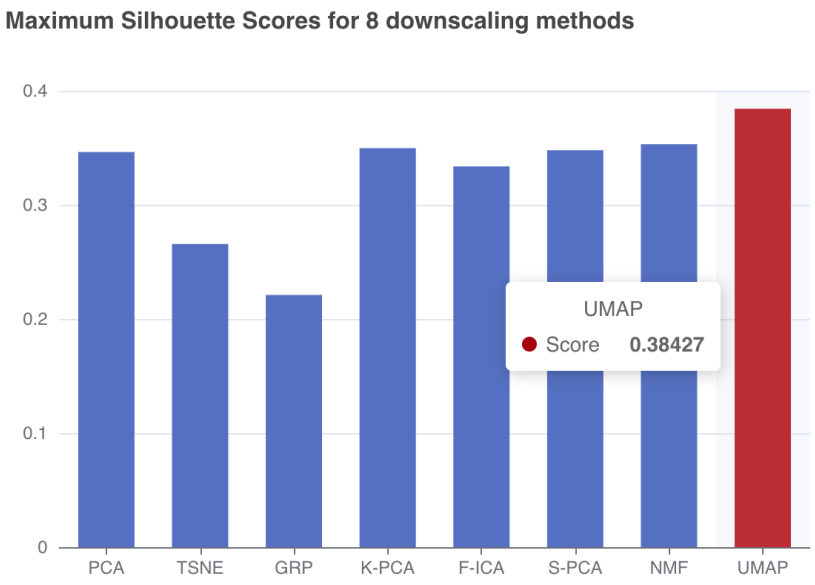

Figure 2: Maximum Silhouette Scores for 8 methods

reduction also facilitates visualization and understanding of the distribution patterns of the data, as demonstrated by the results of the various methods.

The eight dimension reduction methods we employed were Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), Gaussian Random Projection, Kernel PCA, Fast Independent Component Analysis (Fast ICA), Sparse PCA, Non-negative Matrix Factorization (NMF), and Uniform Manifold Approximation and Projection (UMAP).

To determine the optimal number of clusters (k) and the most effective approach for clustering, we employed silhouette coefficients. Silhouette coefficients close to +1 indicate that a sample is significantly distant from neighboring clusters. The heatmap in Figure 1 displays the Silhouette scores for each method and number of clusters, with the largest score of 0.38427 occurring at k=7 under the UMAP method.

4.2 Comparison of different approaches to dimension reduction with k=7

To gain a better understanding and evaluation of the dimensionality reduction approaches, we selected k=7 and conducted silhouette analysis for each of the eight methods. We plotted the silhouette plot for the various clusters on the left, and the visualization of the clustered data on the right. In the silhouette plot, the y-axis represents the clusters (in different colors) with the length encoding the

number of instances in each corresponding cluster. The x-axis represents the silhouette coefficient values, with the red dashed line indicating the silhouette coefficient value for the complete data set.

In terms of the proportion of clusters, most methods exhibit a proper ratio (albeit with some variation) for most clusters, with two exceptions. In tSNE 3(d), cluster 6 takes up only a particularly small proportion, while in kernel PCA 3(b), cluster 2 accounts for a considerably large proportion.

Regarding the spatial distribution of the data, most methods disperse the original data into a 2-dimensional plane, except for tSNE, which maps the first 6 clusters with smaller distances and the last cluster far away from them. We also observe that UMAP, which exhibits the largest silhouette coefficient value, is the most effective method for dispersing the data while minimizing the overlap between different clusters.

4.3 The trend of topics

Selecting the reduction approach (UMAP) and k=7, we can analyze the trend of the conference over the years. Firstly, we determined the topic of each cluster based on the word frequency of the papers it contains 4. Secondly, we divided 29 years into 6 segments: 1987-1991, 1992-1996, 1997-2001, 2002-2006, 2007-2011, and 2012-2015. Thirdly, we plotted the Sankey chart 5, and thus we can analyze the trend of topics over time. For each topic cluster, we summarized its main topic keyword as follows.

By selecting the reduction approach and k=7, we were able to analyze the trend of the conference over the years. Firstly, we determined the topic of each cluster based on the word frequency of the papers it contained 4. Secondly, we divided the 29 years into six segments: 1987-1991, 1992-1996, 1997-2001, 2002-2006, 2007-2011, and 2012-2015. Thirdly, we plotted a Sankey chart 5 to visualize the flow of topics over time. For each topic cluster, we summarized its main topic keyword as follows.

- Cluster 0: Neural network
- Cluster 1: Matrix algorithm
- Cluster 2: Statistical machine learning
- Cluster 3: Machine learning theory
- Cluster 4: Reinforce learning
- Cluster 5: Computer vision
- Cluster 6: Neural Model

The Sankey chart is a powerful tool for visualizing complex data flows and relationships. In the chart 5, the size of the flow between segments of years (nodes on the left) and topics (nodes on the right) represents the number of papers. By using a visual representation of data flows, the chart simplifies the complex data and can help to identify patterns and trends or topics that may not be immediately apparent in raw data.

Our analysis revealed that, over time, the number of papers related to topics 0(Neural network) and 6(Neural Model) gradually decreased, while the number of papers related to topics 1(Matrix algorithm), 2(Statistical machine learning), and 3(Machine learning theory) increased. Topics 0(Neural network) and 6(Neural Model) accounted for over 90% of the total proportion in 1987-1991 and over 70% in 1992-1996. In contrast, topics 1(Matrix algorithm), 2(Statistical machine learning), and 3(Machine learning theory) accounted for over 60% of the proportion in 2007-2011 and over 70% of the proportion in 2012-2015. These findings suggest an increasing trend in research involving machine learning theory and statistical machine learning, while the research trend toward neural networks is decreasing.

5 Discussion

In summary, our approach exhibits several strengths in terms of method comparison, visualization, and interactive exploration. However, it also has some weaknesses regarding the discussion of dimensionality reduction and the limited validation of alternative clustering techniques. Additionally, our study highlights the need to consider word correlations to improve topic extraction results.

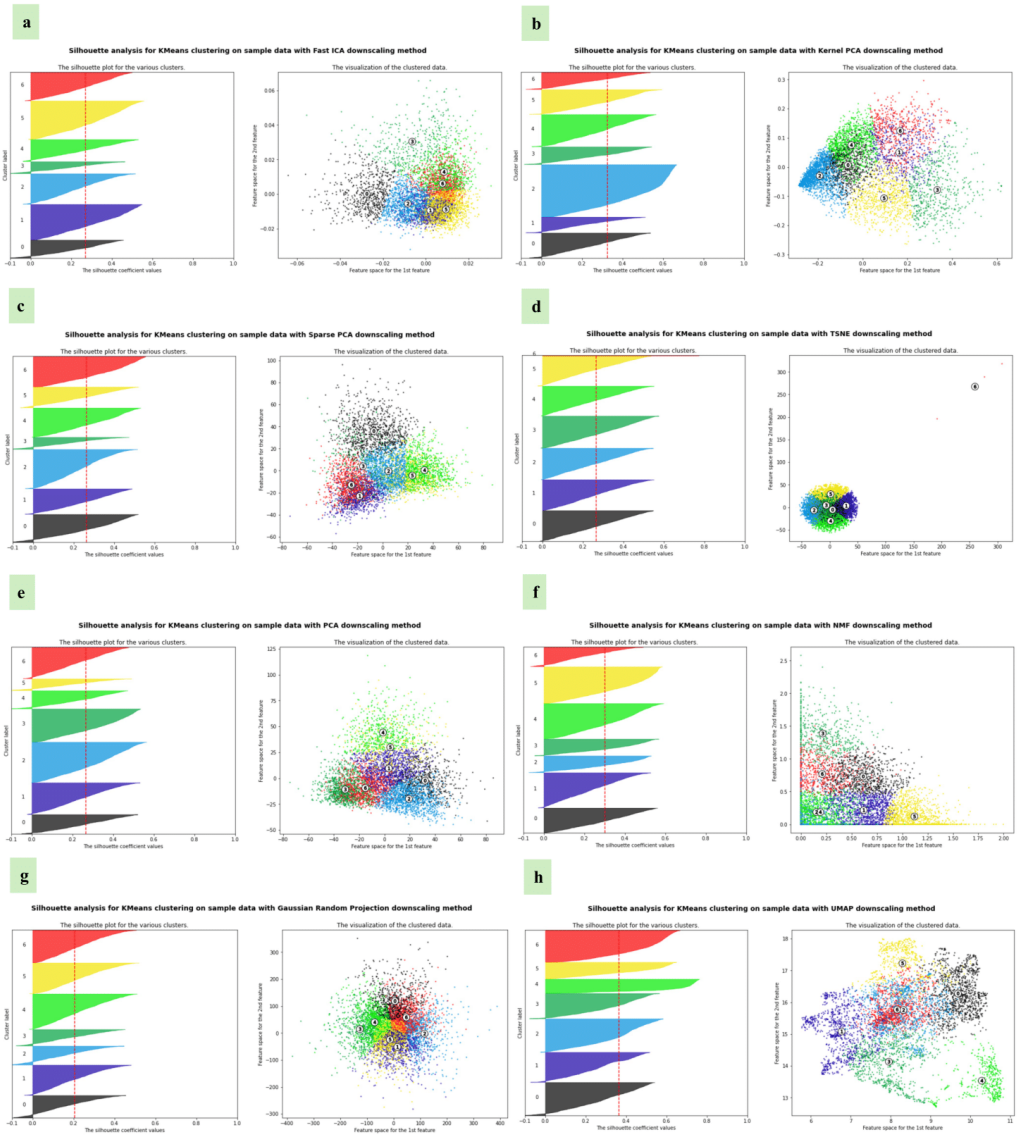

Figure 3: Silhouette analysis on K-means clustering for 8 different dimension reduction methods.

5.1 Strengths

Our study offers several strengths that contribute to the effectiveness of our approach. Firstly, we conducted a comprehensive comparison of various methods, allowing us to gain insights into the superiority of our chosen methodology. Secondly, we employed a visual analytics system that provides an intuitive understanding of the results. The interactive nature of our system enables users to explore the extracted topics and delve deeper into the underlying patterns. This interactive capability enhances user experience and facilitates knowledge discovery.

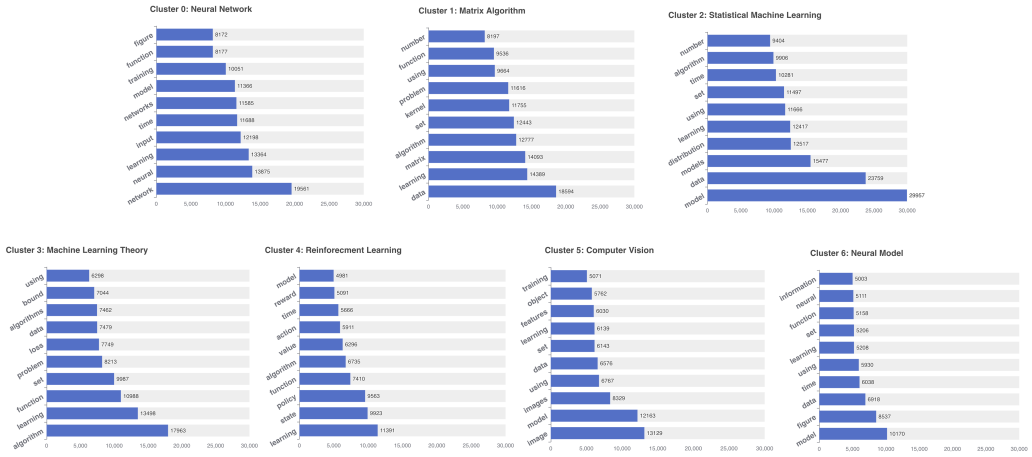

Figure 4: Top 10 frequent words in each cluster

5.2 Weaknesses

While our approach has shown promise, it is important to acknowledge its weaknesses. Firstly, we did not extensively discuss the dimensionality reduction techniques employed in our pipeline. We reduced the dimensions to three for intuitive visualization, but users can adjust this based on their specific requirements. Additionally, we focused on comparing different dimensionality reduction methods and did not extensively validate alternative clustering techniques. Users are encouraged to explore alternative clustering approaches, such as hierarchical clustering, depending on their specific needs.

5.3 Limitations and future work

One limitation of our study is that we solely focused on word frequency as a feature. We did not consider the semantic correlations between different words, such as words with similar meanings. Accounting for such correlations and incorporating measures to handle them could potentially enhance the quality of the extracted topics. Future research should explore methods that take into account the semantic relationships between words to capture more nuanced topic information. Furthermore, we plan to explore the effectiveness of our approach on larger datasets and investigate its potential applications in other domains, such as social media analysis and business intelligence.

References

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

SP Lloyd. Least square quantization in pcm. bell telephone laboratories paper. published in journal much later: Lloyd, sp: Least squares quantization in pcm. *IEEE Trans. Inform. Theor.(1957/1982)*, 18(11), 1957.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

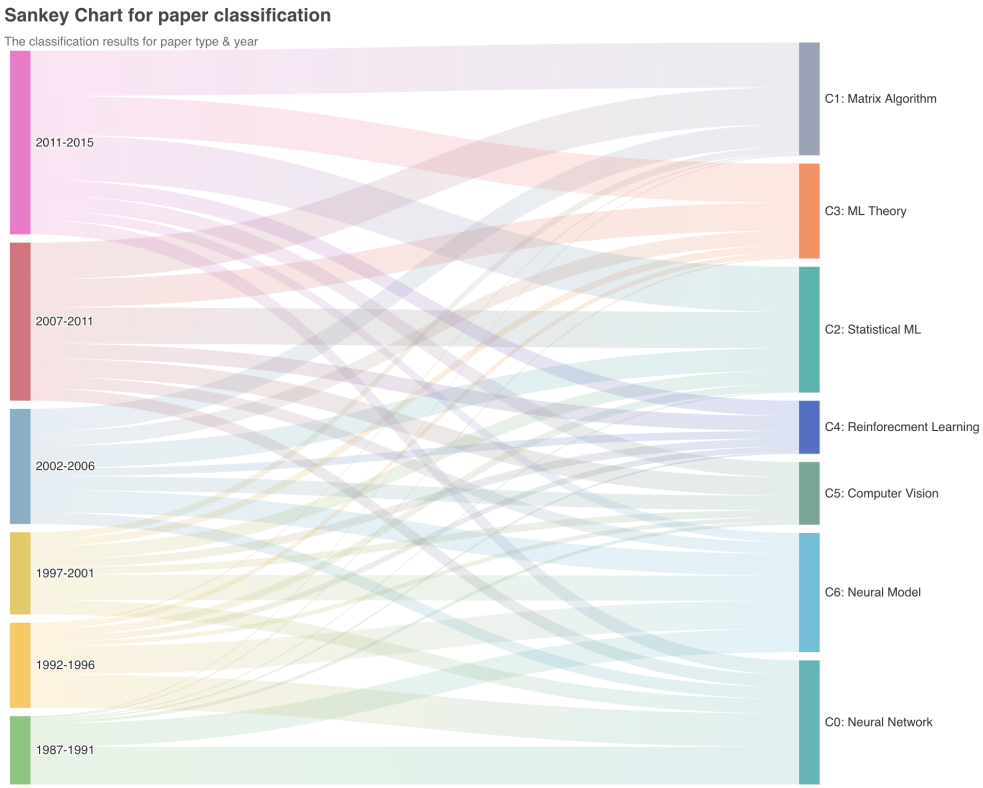

Figure 5: Sankey Chart for the paper classification

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Artificial Neural Networks—ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings*, pages 583–588. Springer, 2005.

Aapo Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110534, 2013.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.