Dimensionality Reduction and Random Projections

Matei Cananau
EECS at KTH
cananau@kth.se

Arvid Ljung
EECS at KTH
arvidlju@kth.se

Gustav Sjökvist
EECS at KTH
gsjokv@kth.se

Project Group 5
December 22, 2024

Abstract

This project implemented the methods discussed in a 2001 paper on Random Projection (RP) for dimensionality reduction. Random projections were applied in both the dense and sparse forms. The performance was compared to other, more or less popular, methods such as Principal Component Analysis (PCA) and Discrete Cosine Transform (DCT). Results showed that PCa consistantly had the lowest reconstruction error, but at the cost of being the most computationally expensive. DCT, primarily used for image and audio data, showed a smooth curve steadily approaching zero error, but was less effective on text data. Dense Random Projection was found to be a good balance between computational efficiency and accuracy, while implementation choices influenced the performance of Sparse Random Projection greatly. Results show that the original paper's findings were mostly reproducible, but with some deviations due to their implementation.

1 Introduction

The work of Bingham and Mannila (2001) presents Random Projection (RP) as a dimensionality reduction method for both image and text data, comparing its performance with traditional techniques such as Principal Component Analysis (PCA) and Discrete Cosine Transform (DCT) [1]. The paper makes a strong case for RP, showing both the theoretical guarantees from the Johnson-Lindenstrauss lemma and the practical advantages to the other techniques in computantional efficiency.

For image data, the study utilized 1 000 grayscale images windows of size 50x50 pixels randomly retrieved from 13 natural scenery images. Each window was flattened into a one-dimensional vector, and the reconstruction error was measured by a calculation of the difference between the original and reduced representations. RP was evaluated alongside PCA and DCT in terms of their ability to preserve data fidelity while reducing dimensions. Sparse Random Projection (SRP) was also introduced as a variation of RP to further improve computational efficiency by utilizing sparsity in the projection matrix.

In the case of text data, the authors selected documents from four specific university newsgroups, and represented them as term frequency vectors without stemming, zero-mean normalization, or variance adjustment. The similarity between randomly chosen document pairs using inner products was evaluated. The error was determined as the difference between pre- and post-reduction values, highlighting each method's capacity to perserve the pairwise similarities in the data.

The study concluded that RP was an efficient and computationally inexpensive method, introducing only minimal data distortion. It demonstrated that RP could produce competitive results with PCA and DCT in preserving data structure while significantly reducing computational costs. This made RP, and especially SRP, a viable alternative to more popular, but more computationally expensive, dimensionality reduction techniques.

In our project, we replicated and expanded upon the findings of Bingham and Mannila by implementing the RP, SRP, PCA, and DCT methods from scratch in python. These were applied to both the original datasets and larger, additional datasets spanning image, text, and other domains. We meant to assess the reproducibility of the original results and examine the modern applicability of these dimensionality reduction techniques to new and diverse data.

2 Methods

We implemented the dimensionality reduction methods discussed in the original paper. Instead of directly importing these methods from external libraries, we implemented them in Python using the NumPy library for matrix operations.

The paper's original datasets were found using its provided links and references. Images obtained from the links had a resolution of 512x256 either in landscape or portrait orientation. The text data was made up of emails between university group members form the 1990s.

In order to replicate these findings, we familiarized ourselves with and implemented following the methods introduced in the paper:

2.1 Random Projection (RP)

Random Projection is an intuitive and straightforward method for dimensionality reduction [2]. It projects data onto a randomly generated lower-dimensional subspace. The random subspace is constructed by sampling a matrix of random values from a Gaussian distribution:

$$Y = X \cdot R, \quad R_{ij} \sim \mathcal{N}(0, \frac{1}{k})$$

where k represents the target reduced dimensionality.

The method is computationally efficient since it avoids the need to calculate eigenvectors and eigenvalues, which may be computationally costly for large datasets.

2.2 Sparse Random Projection (SRP)

Sparse Random Projection (SRP) is a variation of the Random Projection method, where the random subspace is generated using a sparse matrix. Instead of constructing a fully dense matrix with random

values from a Gaussian distribution, SRP introduces sparsity by setting a significant portion of the matrix elements to zero. This sparsity reduces the computational cost of matrix multiplication, which is particularly beneficial for high-dimensional datasets.

The sparse projection matrix R is constructed using Achlioptas' distribution:

$$R_{ij} = \begin{cases} +\sqrt{3/k}, & \text{with probability } \frac{1}{6}, \\ 0, & \text{with probability } \frac{2}{3}, \\ -\sqrt{3/k}, & \text{with probability } \frac{1}{6}. \end{cases}$$

where k is the target reduced dimensionality, and the scaling factor $\sqrt{3/k}$ ensures that the projected data retains its variance properties.

While SRP is designed to provide a computationally efficient alternative to dense Random Projection, the level of sparsity is what actually determines how much faster it will be, and most therefore be chosen carefully. Higher sparsity (more zero elements) leads to faster computation but may slightly degrade the accuracy of the projection. By contrast, lower sparsity increases accuracy but at the expense of computation time. In our experiments, we used a density of 0.1.

The theoretical foundation of SRP comes from the Johnson-Lindenstrauss lemma, which guarantees that distances in high-dimensional spaces can be approximately preserved even after projection onto a lower-dimensional sparse subspace.

2.3 Principal Component Analysis (PCA)

Principal Component Analysis is a widely used technique for dimensionality reduction. It identifies the principal components of the data, which correspond to the eigenvectors of its covariance matrix, representing directions of maximum variance. Although PCA is highly effective, it can be computationally expensive for large datasets. However, approximations help to alleviate this issue. Thus, for the text datasets and the numerical dataset SVD will be used instead, which is a less computationally expensive version of PCA [1].

2.4 Discrete Cosine Transform (DCT)

The Discrete Cosine Transform reduces dimensionality by transforming data into the frequency domain. The theory behind it is that high-frequency components can be discarded without significant loss of information. This philosophy makes DCT especially suitable for image and audio data. The DCT is also used in JPEG compression, where it creates a compact representation of the image data [3].

$$X_k = \alpha(k) \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi k(2n+1)}{2N}\right)$$

3 Results

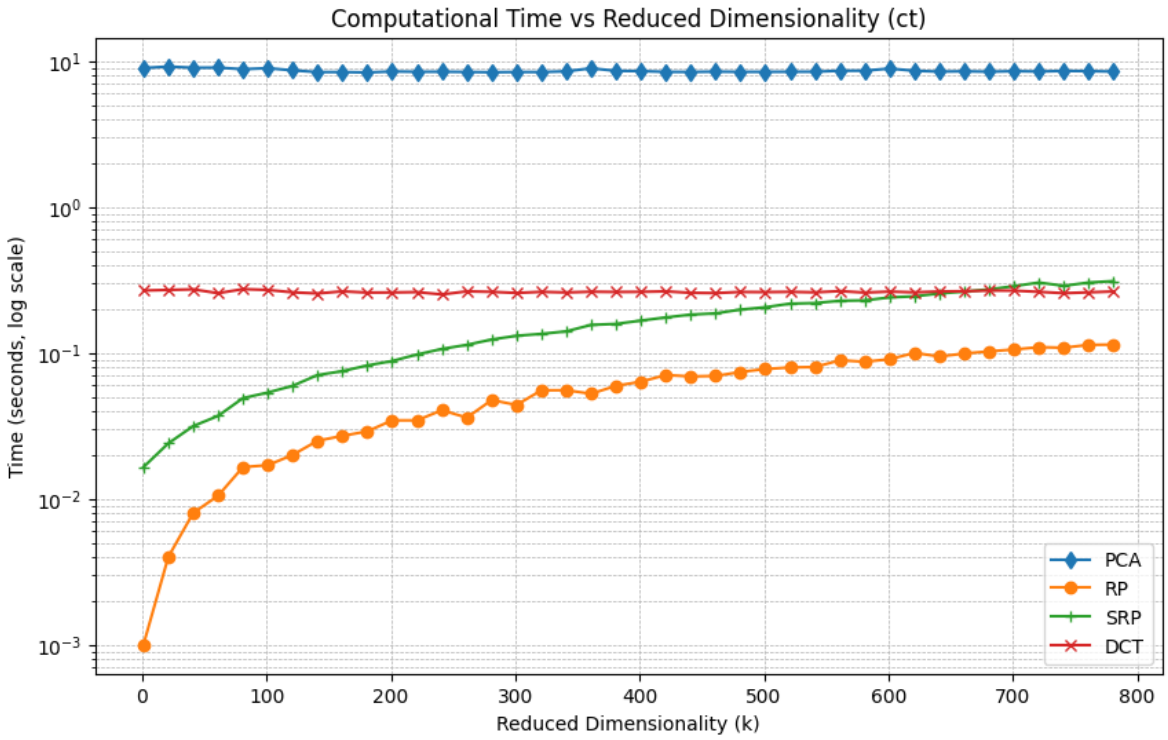

Figure 1: The plot shows the runtime for the CT image dataset for the different algorithms.

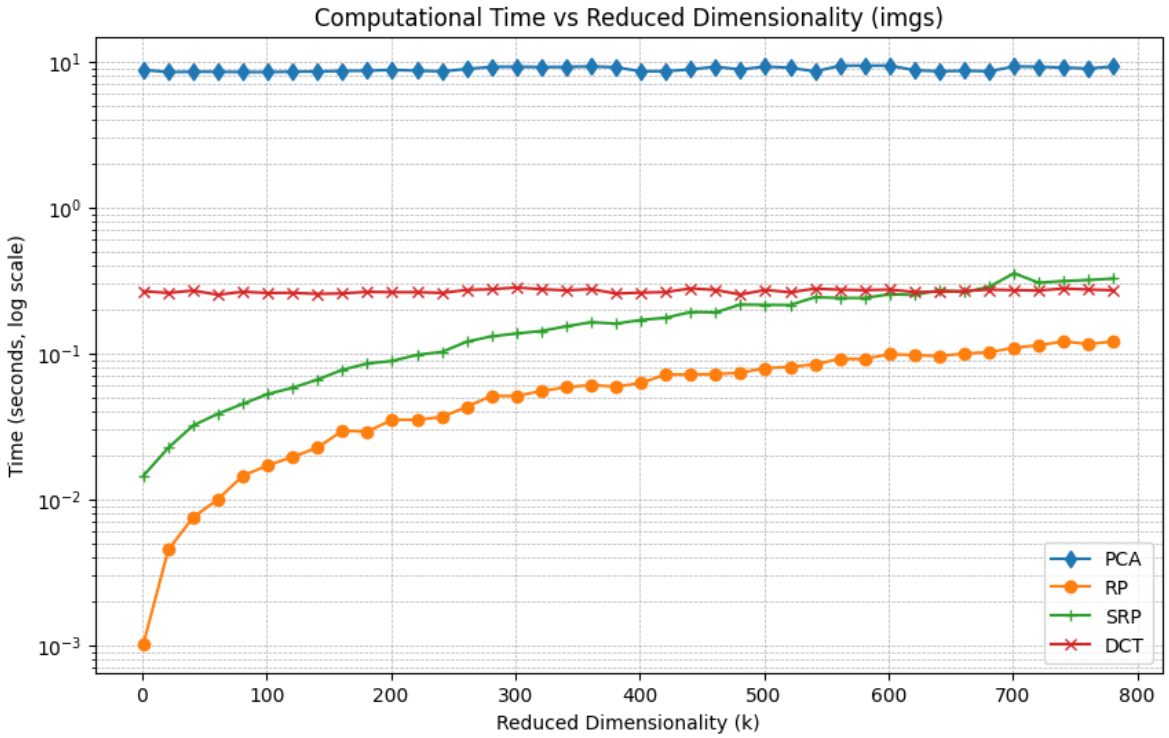

Figure 2: The plot shows the runtime for the original image dataset for the different algorithms.

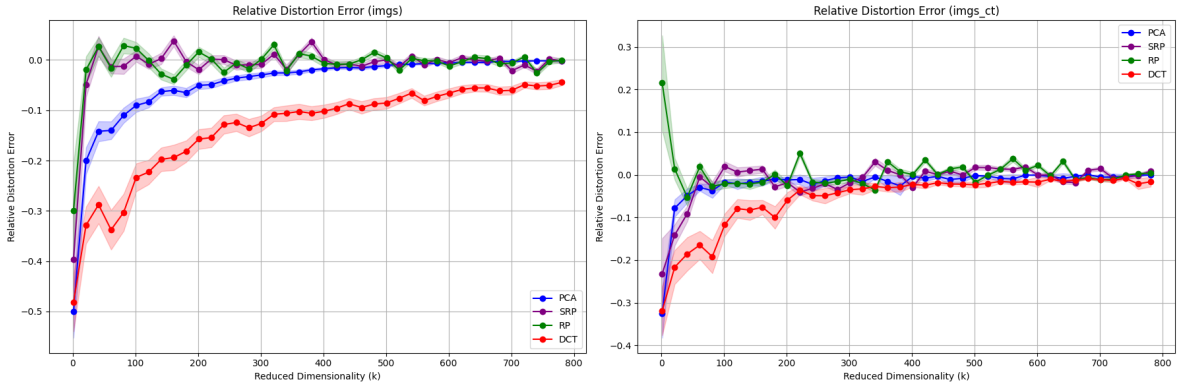

Figure 3: The plot shows the relative distortion error as a function of reduced dimensions (k) for the original image dataset and the CT image dataset using different dimensionality reduction methods.

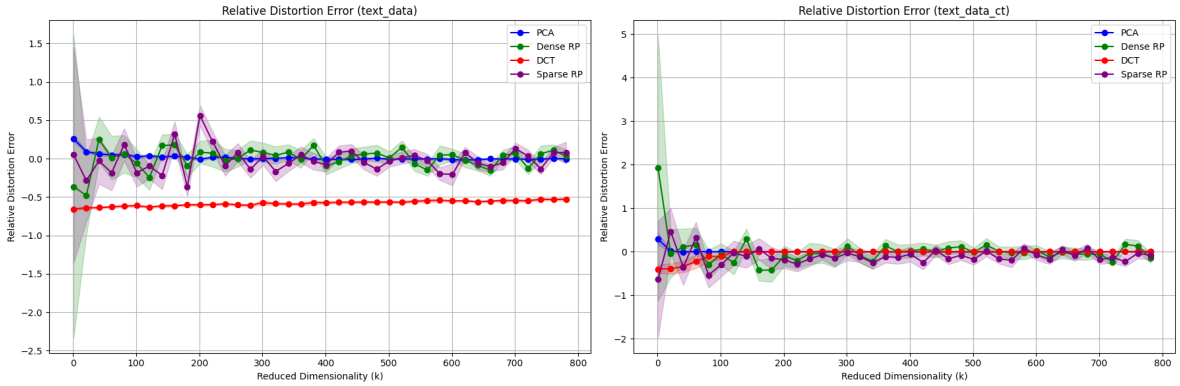

Figure 4: The plot shows the relative distortion error as a function of reduced dimensions (k) for the original text dataset and the CT image dataset using different dimensionality reduction methods.

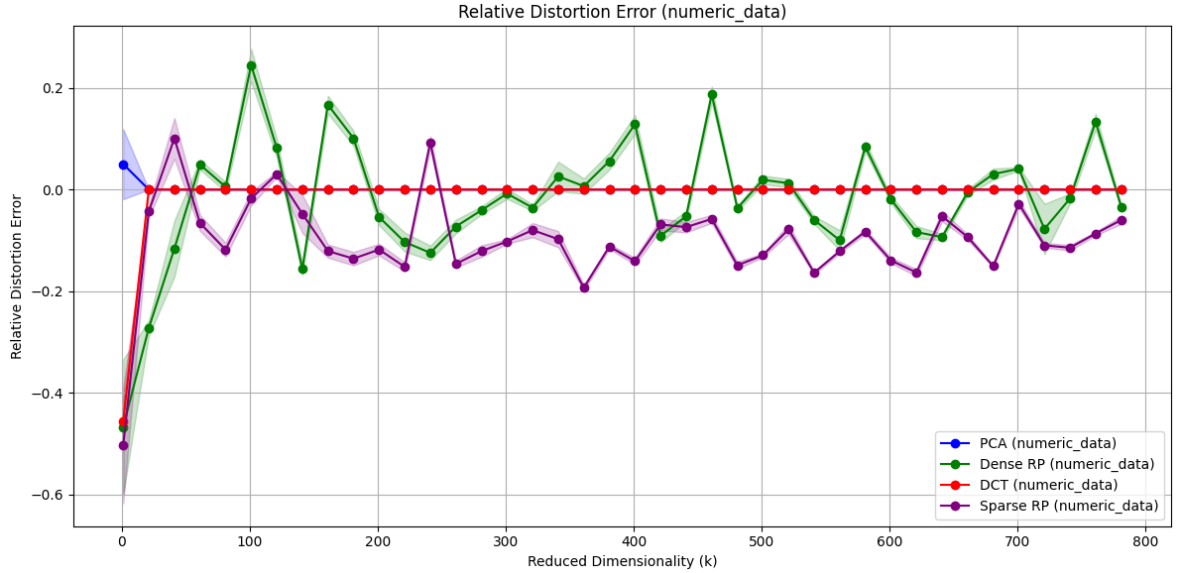

Figure 5: The plot shows the relative distortion error as a function of reduced dimensions (k) for the numeric dataset using different dimensionality reduction methods.

4 Discussion

The resulting graphs were of similar shape to those in the original paper. However, the performance varied slightly between the two implementations.

4.1 Image data

For the original image datasets (Figure 3), the error stabilizes beyond 500 dimensions for PCA. SRP and RP do not stabilize completely for any dimensions but it looks like they are nearly stabilizing for very high dimensions. At higher dimensions, the methods converge in their ability to preserve relative distances. PCA, RP and SRP consistently demonstrate lower errors, with values near zero indicating better distance preservation. This aligns with previous findings in the paper, that PCA is optimized for variance preservation, and RP and SRP approximate pairwise distances effectively as dimensionality increases [1]. PCA's performance consistently good across increasing dimensions, with a significant reduction of error at low dimensions. This shows its ability in preserving the structure of noiseless

image data after dimensionality reduction. RP shows a similar trend, with a big decline in error at lower dimensions and stabilization at higher dimensions aligning closely with PCA's performance. SRP shows similar behavior to Random Projection and shows a high decline in error for higher dimensions. DCT on the other hand shows consistently higher errors than the different algorithms for all dimensions and does not get a value near zero, which is consistent with the findings of the paper [1]. DCT shows consistently higher errors compared to PCA, SRP, and RP, even more so at higher dimensions. This suggests that DCT is less effective in maintaining relative distance preservation for noiseless image datasets. Confidence intervals reveal greater variability in error at lower dimensions for methods such as RP and DCT which becomes significantly smaller beyond 100 dimensions. This indicates increased reliability of results as dimensionality increases.

For the CT image dataset, the error stabilizes around 100 dimensions for RP, SRP, and PCA, slightly earlier than in noiseless datasets. This suggests higher compressibility or simpler structure in the CT images. The performance of the algorithms is different from the original dataset. Here all the algorithms show low errors (values near zero) for higher dimensions. DCT especially seems to perform much more on the new dataset compared to the other. Confidence intervals narrow significantly beyond 100 dimensions indicating more reliable results at higher dimensions, compared to the lower dimensions where the confidence intervals display higher uncertainty.

4.2 Text data

For the text datasets (as seen in Figure 4), PCA maintains consistently low error rates across all dimensions, while RP and SRP stabilize as dimensionality increases. PCA's rapid error stabilization near zero mirrors the results in the literature which demonstrates its effectiveness in preserving data structure during dimensionality reduction [1]. RP stabilizes around zero error at smaller dimensions, consistent with findings in prior studies [1]. SRP and RP, while stabilizing quickly, displays slightly higher errors than PCA. Even though random projection (RP and SRP) preserves distances quite well. DCT on the other hand performs poorly across all dimensions similar to its its weaker performance on image data. The shaded confidence intervals show greater variability at lower dimensions for all methods, becoming narrower as dimensionality increases showing increased reliability on the results for the higher dimensions.

For the CT text dataset, the error rate stabilizes earlier for all methods at around 100 dimensions with error values near zero, while PCA stabilizes quicker with consistently low errors. DCT stabilizes later than the other methods but performs better than in the other text dataset, with errors near zero for high dimensions. DCT's strong performance at higher dimensions shows its potential for text data applications, despite its weaker performance in image data and other text datasets. Overall, the earlier stabilization in the CT text dataset indicates that dimensionality reduction methods can preserve distances more effectively in this dataset with fewer dimensions. These observations indicate that the results of the paper generalize effectively to the CT text datasets which shows the robustness of PCA and RP performing on new datasets [1].

4.3 Numeric data

In the numeric dataset, the methods show some differences in their behavior (as in Figure 5). PCA and DCT consistently maintain low errors across all dimensions, with DCT maintaining a nearly constant error close to zero indicating exceptional stability, perhaps too good to be true. PCA while slightly more variable, also stabilizes quickly at low error values. According to these results, these methods seem highly reliable for numeric data reduction. RP and SRP show more significant variability, particularly at lower dimensions. RP stabilizes more quickly than SRP and approaches errors close to zero at higher dimensions. However, SRP demonstrates higher errors and higher variability across all dimensions which suggests that it is less reliable for numeric datasets compared to RP, PCA, and DCT. Confidence intervals reveal greater variability for SRP and RP at lower dimensions indicating less consistent performance in this range. These intervals narrow significantly for RP and SRP at higher dimensions, indicating increased reliability as dimensionality increases. These results suggest that PCA and DCT are the most suitable methods for high-accuracy tasks involving numeric data, given their stability and low error rates. RP and SRP remains a viable alternative for scenarios prioritizing computational efficiency, while Random Projection's higher variability makes them less accurate, but more computationally efficient [1].

4.4 Performance

Our results (demonstrated in Figures 1 and 2) deviate slightly to that of the original paper in terms of the performance of the algorithms. In both cases, PCA was very accurate, but at the cost of performance. The complexity, $O(nd^2 + d^3)$, given a dataset X of size $n \times d$, also support this.

DCT was significantly faster than PCA, with almost two orders of magnitude which is supported by the complexity of $O(nd \log d)$. The runtime did not change with larger values of k, which was expected.

Our results showed that Dense, or regular, RP consistently outperformed Sparse RP and was the fastest among the four methods. This was in contrast to the original paper, where SRP was the fastest. The complexity $O(ndk)$ of RP scaled well with increasing dimensionality reduction.

SRP, however, was expected to be the fastest of the four as was the case in the original paper, given it's complexity of $O(ndk * s)$ where s is the sparsity level, or density. This discrepancy might be attributed to the method of implementation, including the sparsity level and the overhead of constructing the sparse projection matrix. There were some issues with memory usage in earlier iterations of our implementations, so we opted to construct the sparse projection matrix row-by-row instead of generating it in one operation. This design choice allowed us to handle large datasets without requiring excessive memory, as the matrix components were stored in a compressed sparse row (CSR) format. However, this approach introduced significant overhead due to repeated random sampling and iterative assembly of the sparse matrix. While this resulted in slower runtime compared to a dense random projection, it ensured that the method remained feasible for high-dimensional datasets with limited computational resources. We believe the results might have been different with a different implementation.

5 Learning Summary

We have now seen how dimensionality reduction techniques function by implementing several algorithms and comparing their outcomes. This process made the difference between these methods clear, including their advantages and limitations. A difficult decision to be made by those new to the field of machine learning and data science as a whole, is the trade-off between achieving precise results yet requiring a lot of computational resources.

For example, Principal Component Analysis often reached the lowest reconstruction error out of all the tested methods, but at the cost of being computationally expensive, limiting its practicality for larger datasets. In contrast, methods such as Dense Random Projection and Sparse Random Projection were much faster to compute and demanded fewer resources, but produced less accurate results.

Furthermore, we learned the importance of outlining the structure and roadmap of any data-related project. Early on, we tried implementing the algorithms without a concrete philosophy, as every group member had their own idea of how to proceed and even how to read in and preprocess the data. This led to confusion and inefficiency. Realizing that the algorithms come last, after the data has been properly cleaned and prepared for use, we change our approach and outlined how we wanted to format the data, and then wrote the algorithms to parse it. We believe dealing with the datasets served as the key insight, realizing that it is much harder to procure, clean and format good data than it is to simply implement an algorithm that has already been invented.

Contributions

Each member of Project Group 5 contributed equally to the project. It can be found using the following link: https://github.com/Matdrox/dimensionality-reduction, and it was worked on in a shared Deepnote notebook. The entire project was collaborative, with all members contributing to all of the implementations and tasks. However, each member took the lead on a specific part of the project:

- **Matei Cananau**: Implementation of DCT, local testing, and article structure
- **Arvid Ljung**: Implementation of PCA and error calculation
- **Gustav Sjökvist**: Implementation of RP and SRP and evaluation of performance

References

[1] 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*. ACM, 2001.

[2] SciKit. Random projection. https://scikit-learn.org/stable/modules/random_projection.html, No date. Accessed: 2024-12-22.

[3] SciPy. scipy.fftpack.dct. https://docs.scipy.org/doc/scipy/reference/generated/scipy.fftpack.dct.html. Accessed: 2024-12-22.