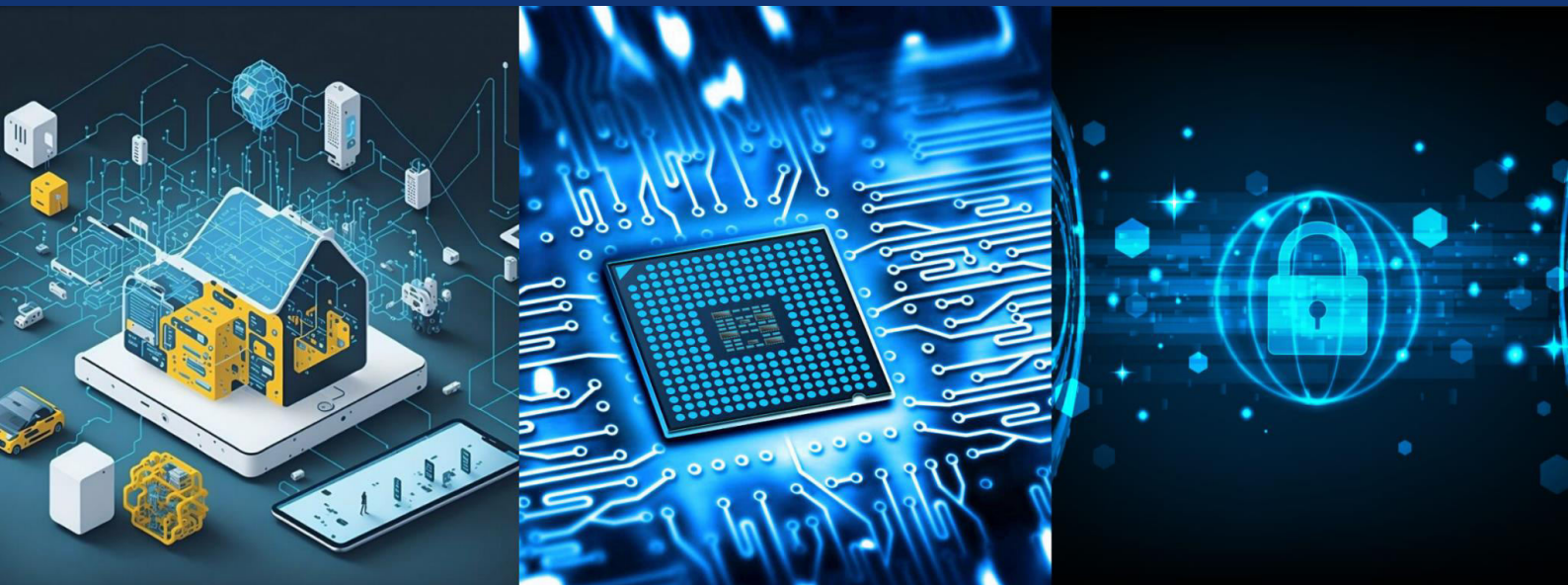
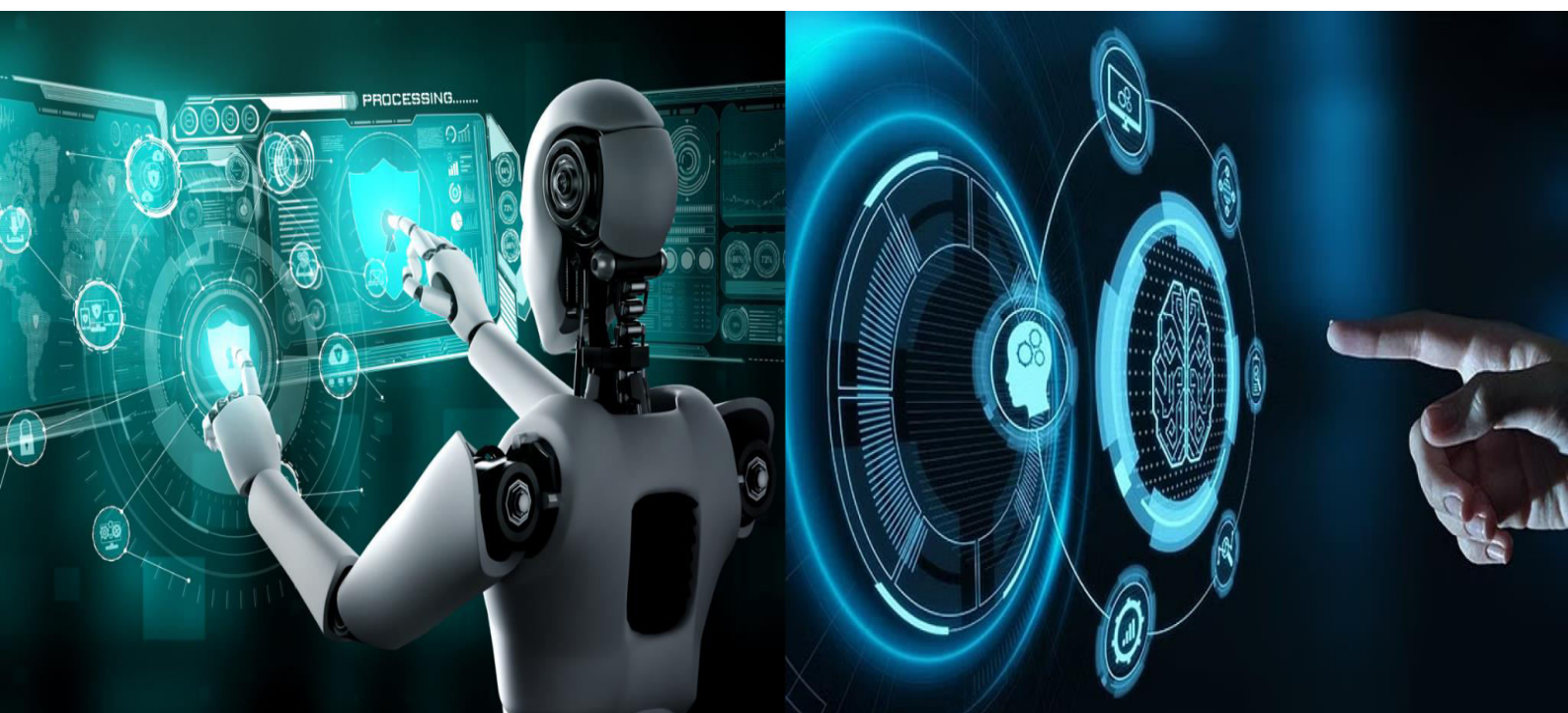




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Why Linear Algebra Is Fundamental to Algorithmic Thinking: A Structured Analysis of Vectors, Transformations, Dimensionality, and Computational Representation

Attharv Bhagchandka

Student, Shri Ram School Moulisari, Gurgaon, India

ABSTRACT: Linear algebra is often said to be the basis of machine learning, but without a clear link to the operations of the algorithms it's hard to grasp. This study explores the impact of vectors, distance measure, matrix decompositions, transformations, and dimensionality on the functioning of a working classification algorithm. The reason for choosing Fashion-MNIST is that every 28×28 grayscale image can be easily and naturally stored as a 784-dimensional vector, and each image has a meaningful class label. Instead of starting with a trained machine-learning model, the analysis started from the underlying binary data, validating and decoding the data, investigating vector geometry, building class centroids, nearest-centroid classification, principal component analysis (PCA) of the covariance matrix, testing low-dimensional representations, manipulating coordinate geometry, and testing the computational costs. The results reveal that the Euclidean classification of centroids is quite accurate, with a 67.68% accuracy on the results using no learned nonlinear features, and that 24 principal components preserve 67.30% of accuracy while discarding 96.94% of dimensions, and that a full orthogonal coordinate transformation achieves 100% accuracy on all 10,000 predictions, and that selectively reweighting one principal direction can change thousands of predictions. The results have shown that linear algebra enables algorithms to be computationally feasible but it also plays a key role in the very nature of the similarity or information that an algorithm considers, as well as the nature of what it considers to be a useful structure for computation.

I. INTRODUCTION

Today's algorithms rarely deal with other objects as humans do. Photographs become an array of numbers, text becomes a vector/matrix, and a series of observations becomes a high-dimensional data structure. Linear algebra provides the language for these representations: vectors represent observations, matrices represent sets of data, and transformations, inner products, norms, and eigendecompositions represent similarity and directions of systematic variation. In this regard, linear algebra, analytic geometry and matrix decomposition are identified as key mathematical tools for machine learning (Deisenroth, Faisal, & Ong, 2020). Principal component analysis is a case in point. PCA is a technique that was originally developed by Pearson (1901) and Hotelling (1933) which maps correlated original coordinates onto orthogonal directions with the greatest variance in each successive direction. The modern treatment focuses on the fact that PCA is a dimensionality-reduction technique as well as an eigenvalue/eigenvector problem (Jolliffe & Cadima, 2016). The matrix-approximation results of Eckart and Young (1936) are also related in a fundamental way to the low dimensional approximation. However, there are some precautions for dimensionality and distance. Common distance measures may yield non-intuitive properties in high-dimensional spaces (Aggarwal, Hinneburg & Keim, 2001), and various similarity definitions may highlight different aspects of the identical observations. By the vector-space tradition in computing, the representation and similarity measures are also a part of how algorithms retrieve or match objects (Salton, Wong, & Yang, 1975). The question of whether it is possible to compress images with PCA is actually too narrow and is extended to how PCA's linear-algebraic representation affects the information, decisions, and computational behavior of an algorithm. Fashion-MNIST is a suitable controlled environment. It presents 70,000 28×28 grayscale fashion images equally distributed over 10 categories: 60000 images in the training set, and 10000 images in the test set (Xiao, Rasul, & Vollgraf, 2017). The analysis is done from the original binary representation to the vector geometry, classification, matrix decomposition, transformation, compression, and computational cost.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

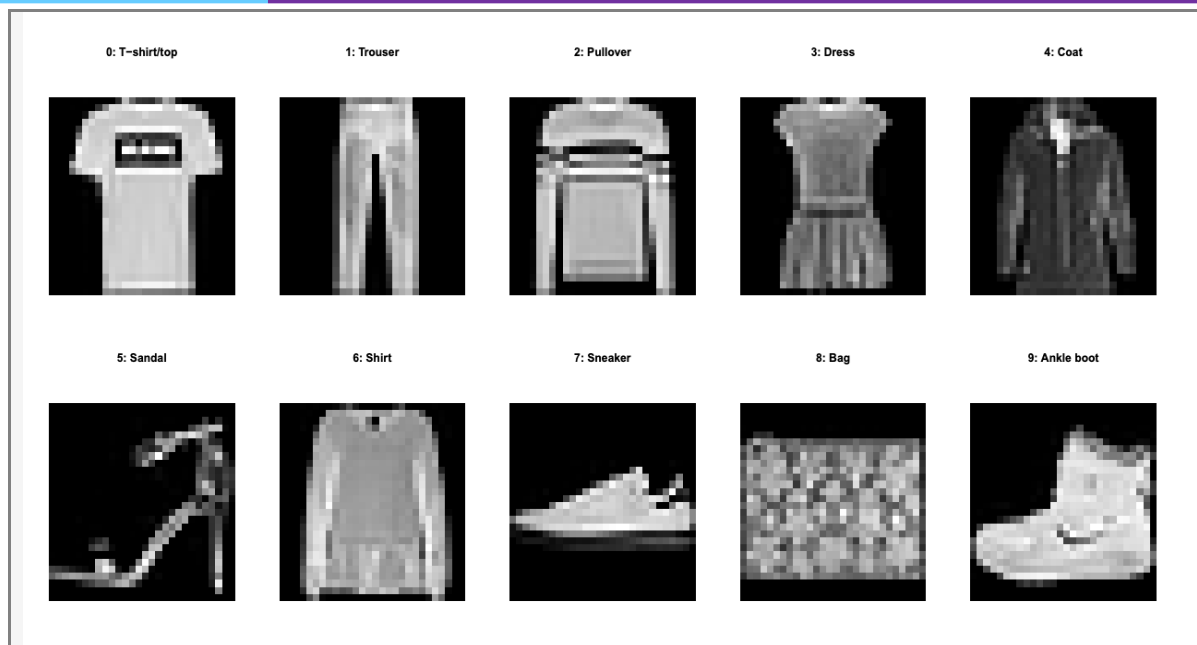


Figure 1. Reconstruction of 784-dimensional vectors into its original 28×28 image structure.

II. DATA AND METHODOLOGY

The methodology was designed to build the data architecture from the data, and not start with a packaged classifier. All the four compressed Fashion-MNIST IDX files were downloaded straight from the source. They were compared to published checksums using their MD5 hashes, and all matched completely. The IDX headers were then manually decoded, and found to contain 60,000 training observations, 10,000 test observations and a magic number of 2051 for an image-file and 2049 for a label-file. The labels were broken down into the unsigned bytes 0-9, and the total number of training observations and test observations per class were 6,000 and 1,000, respectively. These images were then converted to numerical vectors in $\mathbb{R}^{784}$. This produced a $60,000 \times 784$ training matrix containing 47,040,000 pixel values and a $10,000 \times 784$ test matrix containing 7,840,000 values. No missing observations and Pixel intensities varied from 0 to 255. In order to reconstruct the vectors back to 28×28 images, it was necessary to confirm that the order of the vectors had not been lost, just by relying on the number of rows and columns of the matrix. The first step of analysis focused on raw geometry of vectors. L1 and L2 norms, non-zero coordinates and mean intensity were computed for each of the training images. Since the magnitudes of the vectors varied greatly among classes and 50.21% of the coordinates in the training matrix were zero, Euclidean distance and cosine similarity were tested. A paired experiment was then conducted, in which 500 anchors were chosen from each class, for a total of 5,000 anchors, and paired with one of the same class and one of a different class. This paired design was used to determine whether there was a systematic relationship between semantic membership and geometric proximity, which was not compared to pools of unrelated distances. The mean 784 dimensional vector of each class was then taken to be the class centroid. Within-class dispersion and between-class separation were quantified using distances from observations to their own centroid and distances between competing centroids, respectively. The same centroids were used in a deliberately simple nearest centroid classifier, where each of the 10,000 test observations were assigned to the centroid closest in terms of geometric distance. The accuracy, precision, recall, F1 scores, confusion matrices and paired McNemar testing were used to compare the Euclidean and cosine versions. This of course resulted in the matrix-structure stage. Each of the 784 pixel differences was first examined individually, and then the 784×784 covariance matrix was eigendecomposed to get the centered 784 pixel differences. No standardization of pixels, since all the coordinates were already measured on the same intensity scale of 0–255; if low-variance boundary pixels were forced to unit variance, then the image geometry relevant to the reconstruction would be changed. The PCA components were not selected arbitrarily, but from the observed cumulative variance spectrum; they were 1, 3, 10, 24, 84, 187 and 459. For every checkpoint, unseen test images were reconstructed from the mean in the training set and the training-derived eigenvectors, and then compared in pixel space to the reference images with respect to the retained energy, Frobenius



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

error and pixel RMSE. These reduced representations were then fed to the same nearest-centroid classifier, but without any changes. Dimensional representation was the manipulated quantity with the classifier kept fixed. Lastly, distinct mathematical phenomena were separated in the process of transformation and computational experiments. A full 784-dim orthogonal-pca rotation was used for testing geometry (no geometry change) and coordinates (coordinate change). Uniform scaling was conducted to determine if decisions were affected by changing the numerical units. The relevance of an invertible transformation was tested by deliberately scaling only PC1 by factors between 0.25 and 8. Modules E and F then measured storage, arithmetic, empirical run-time, re-use costs, and robustness for metrics and transformations and dimensional levels.

III. RESULTS

3.1 Vector geometry carries class information

There is class information transmitted by 3.1 Vector geometry. There was a large range in the scale of the raw vectors. The mean L2 norm of the training observations was 3,098.81, and the mean L2 norm of the Coat was 3,749.94 while the mean L2 norm of the Sandal was 2,032.72. The average number of non-zero pixels occupied was 508.6 for Pullover, while the average number of non-zero pixels occupied was 251.8 for Sandal. These differences suggested that cosine similarity be used in addition to Euclidean distance, and that absolute magnitude was not necessarily informative. The paired experiment showed a high level of semantic geometry. The mean separation advantage was 734.03 with the mean same-class Euclidean distance of 2,233.23 and a mean between-class Euclidean distance of 2,967.25. In 80.38% of paired comparisons, two observations from the same class were closer, Cohen's $d_z=0.851$. The same conclusion was drawn by using cosine similarity: 0.7532 within class, 0.5767 between classes and 82.28% directional success rate, with $d_z=0.910$. Both paired tests were very significant.

```
pair_design_audit

##  anchors anchors_per_class same_class_comparisons different_class_comparisons
## 1      5000                500                5000                5000

pair_geometry_summary

##           metric same_class_mean different_class_mean same_class_median
## 1 Euclidean distance      2233.2271      2967.2525      2164.3648
## 2 Cosine similarity         0.7532         0.5767         0.7939
##  different_class_median mean_separation_advantage directional_success_pct
## 1              2994.0215              734.0254              80.38
## 2              0.5839              0.1766              82.28
##  cohens_dz
## 1      0.851
## 2      0.910

inferential_summary

##           metric t_statistic  df          p_value ci_lower ci_upper
## 1 Euclidean distance    60.183 4999 <0.0000000000000002  710.1149 757.9359
## 2 Cosine similarity     64.322 4999 <0.0000000000000002   0.1712  0.1819
```

Table 1. Same-class versus different-class vector geometry.

This overlap was also organized. The closest centroid pairs were Pullover–Coat (Euclidean distance 633.09; cosine similarity 0.9841), Pullover–Shirt (682.17; 0.9832), and Coat–Shirt (770.91; 0.9874). Later, the error regions of the classifier were the geometrically overlapping categories. The Euclidean nearest centroid classifier got 67.68% accuracy and 67.25% macro F1 score, and cosine classifier got 66.52% and 64.92% respectively. Shirt recalled 21.7%, Pullover 45.0% and Coat 56.2% in line with overlapping centroids, as is seen with Euclidean distance. In all 45 pairs of classes,



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the centroid distance had a negative correlation with the mutual Euclidean confusion (-0.6917), and the cosine centroid similarity had a positive correlation with cosine confusion (0.8575). The geometry therefore forecasted not only the overall performance but what errors the algorithm hadn't made.

3.2 Full algebraic dimension does not imply full practical dimension

The centered training matrix had a numerical rank of 784 and all of the 784-pixel coordinates had non-zero variance. Hence, dimensional compression cannot be done by dropping the constant variables alone. Instead, the important result emerged from the spectrum: PC1 explained 29.04% of total variance; three components explained 52.81%; ten 71.99%; 24 80.11%; 84 90.06%; 187 95%; and 459 99%. The numerical rank was 784 but its effective rank was merely 29.15.

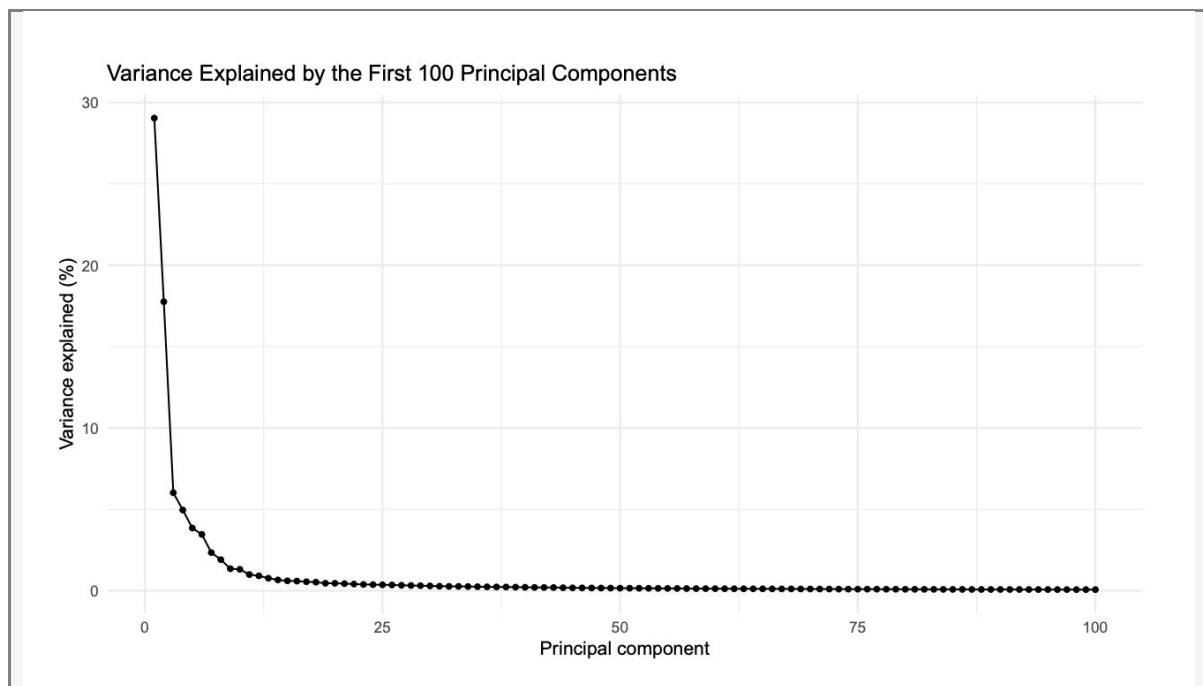


Figure 2. Variance explained by the principal components and percentage of variance accounted for.

This concentration has been generalized to the test-set using test-set reconstruction. There were 84 components with 90.06% of the training variance and 89.99% of test centered energy retained. Pixel RMSE fell from 63.17 at one component to 39.79 at 10, 33.56 at 24, 23.75 at 84, 16.91 at 187, and 7.65 at 459 components.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

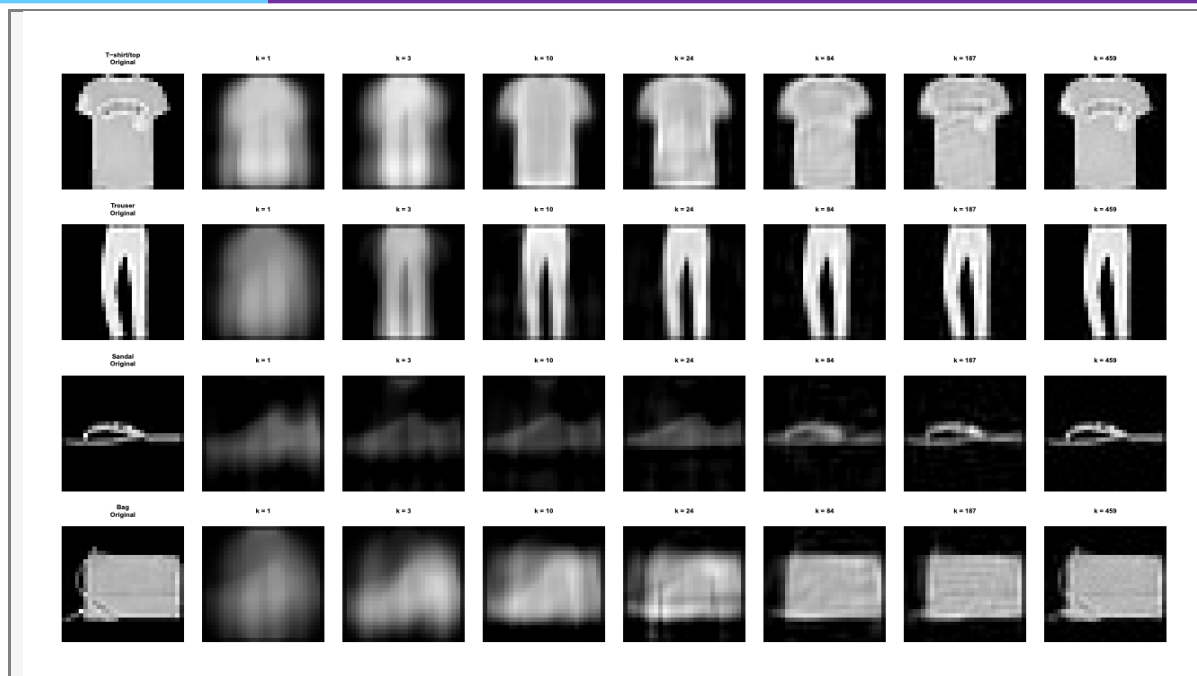


Figure 3. Reconstructing unseen Fashion-MNIST images on increasingly rich PCA representations.

Most important of all, there was significantly less information needed for classification than for the near exact reconstruction. Full 784D accuracy was 67.68%; 187 dimensions produced 67.67%, 84 produced 67.59%, and 24 produced 67.30%. Removing 760 of 784 coordinates (96.94% of original dimensionality) thus only resulted in a loss of 0.38 percentage points in accuracy. Even 10 dimensions achieved 65.66%. Only when the compression was very high, did there result a sharp collapse: 3D accuracy was 51.27% and 1D accuracy 27.56%.

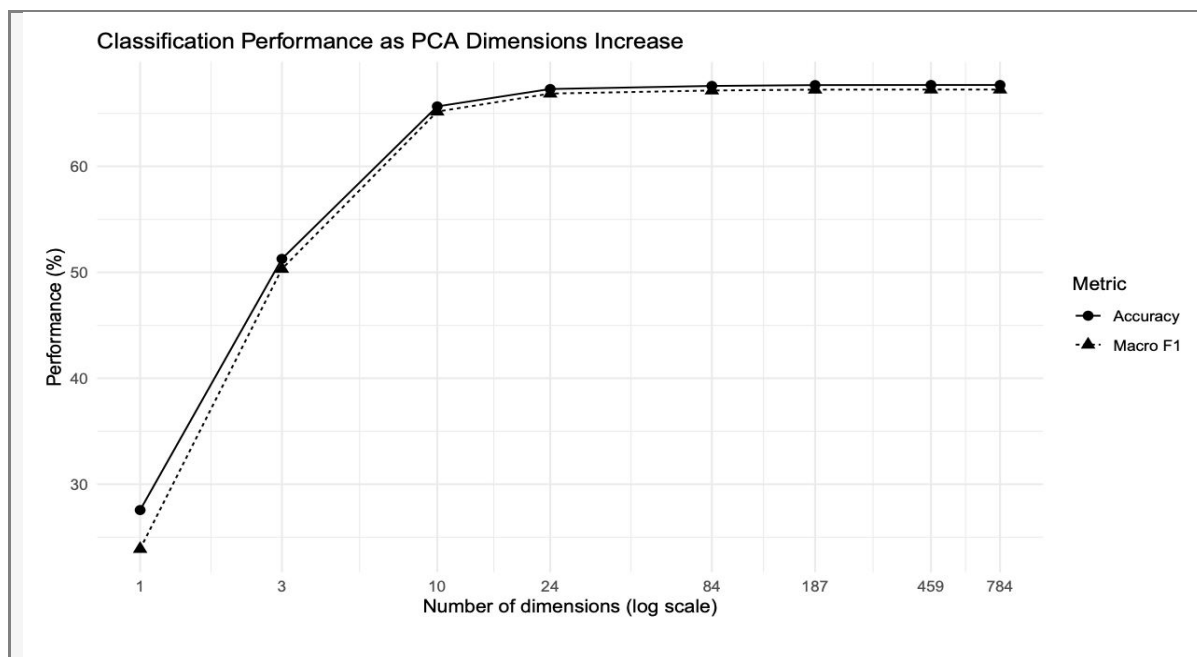


Figure 4. The accuracy of classification and the macro F1 values with respect to the dimensionality of the PCA.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This was further supported by the paired analysis. At 84 dimensions, 99.82% of predictions matched the predictions made by the 784D classifier and the Holm adjusted McNemar test failed to find a conventional 5% difference. The agreement was still 99.20% at 24 dimensions. Every one of the 10,000 predictions was the same as full space with 459 components.

3.3 Representation can change completely without changing the algorithm

The maximum orthogonality error was 6.56×10^{-13} with a full orthogonal transformation transformation. The largest difference in the pair distances allowed was about 1.00×10^{-11} . More importantly, classification was not affected by rotation, it was at 67.68% both before and after rotation, which resulted in 100% agreement in predictions and 0% change in predictions. The coordinates had all changed, but not the Euclidean geometry. If we multiply all of the coordinates by four then we get the same invariance. The algorithmic behaviour was completely changed by direction selective scaling, however. Weights of 0.25 of the original weight increased the accuracy to 70.72%, while doubling and quadrupling the weight reduced the accuracy to 58.95% and 45.24% respectively, and increasing it by 8 times gave an accuracy of only 38.89%. There were 5,135 predictions that changed at $8\times$, and the agreement with the baseline was 48.65%.

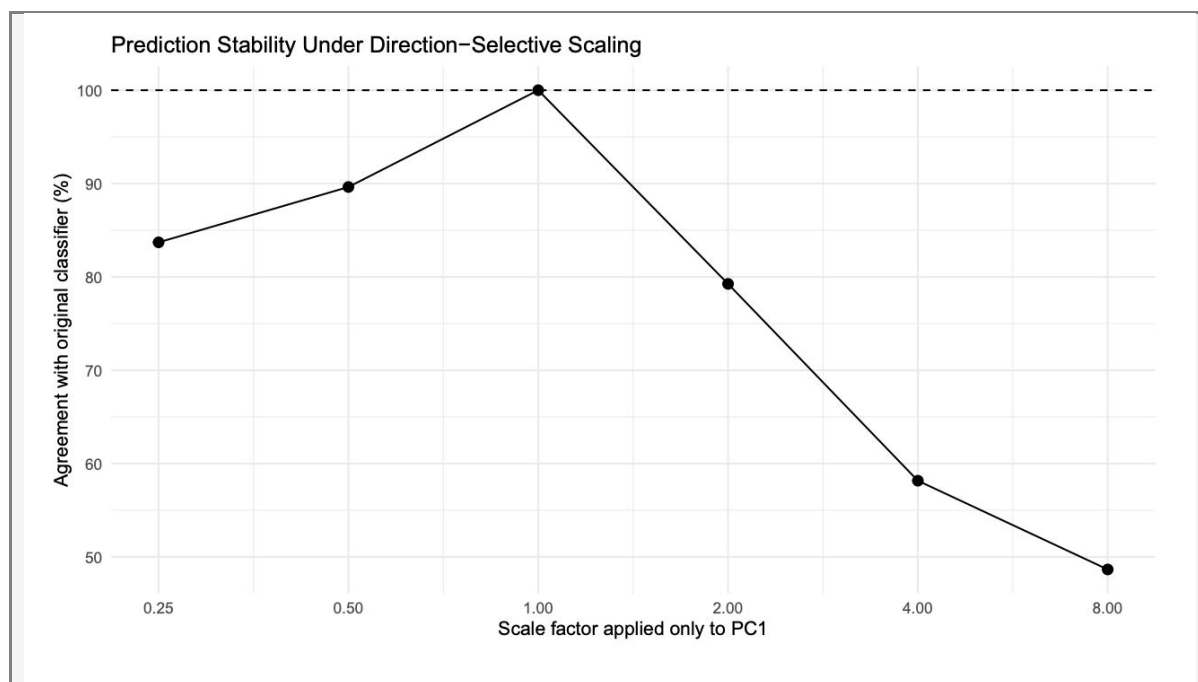


Figure 5. Prediction agreement is only the first principal-component direction which is reweighted..

PC1 was not just noise, as it accounted for 29.04% of the total variance, and 58.16% of the variation in PC1 was between classes, and 41.84% of the variance in PC1 was within classes. Rather, it yielded a coarse class separation, which was not equally beneficial for each boundary. For Shirt, the recall increased by 10.5 percentage points, while recall for Pullover increased by 9.8 percentage points, T-shirt/top by 8.6 percentage points, and by 7.3 percentage points for Dress. That shows that a direction which maximises overall variance does not necessarily merit equal domination in a classification distance function.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

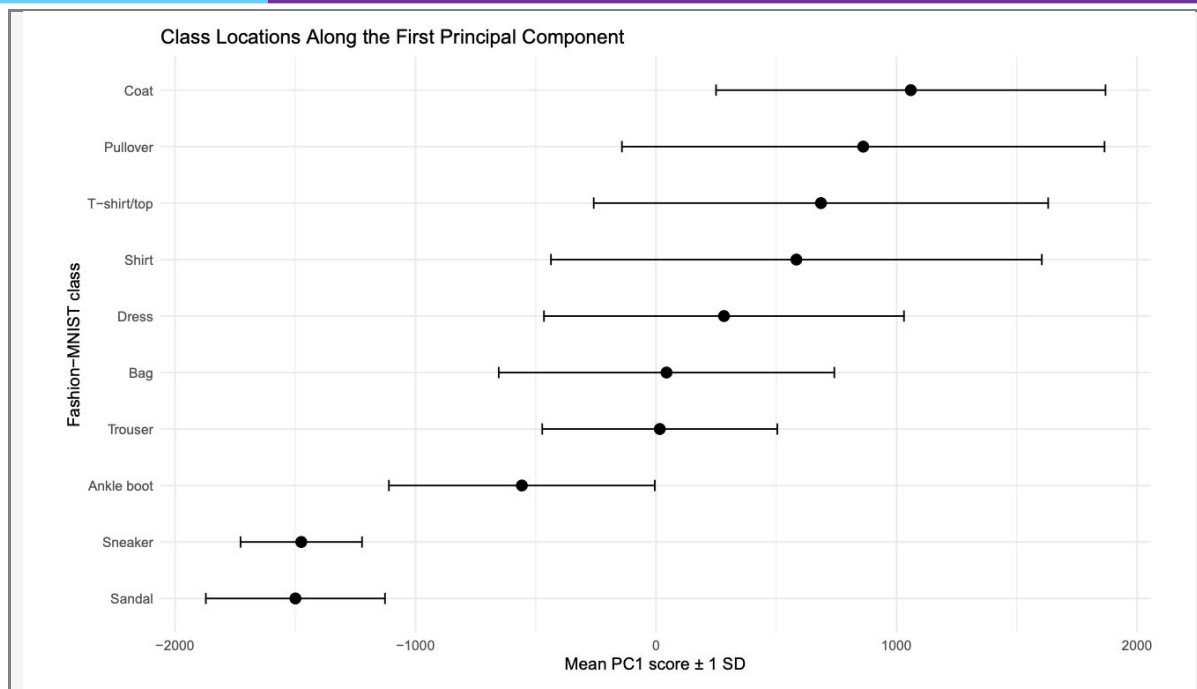


Figure 6. Class locations along first principal component (PC1), mean scores and within-class dispersion.

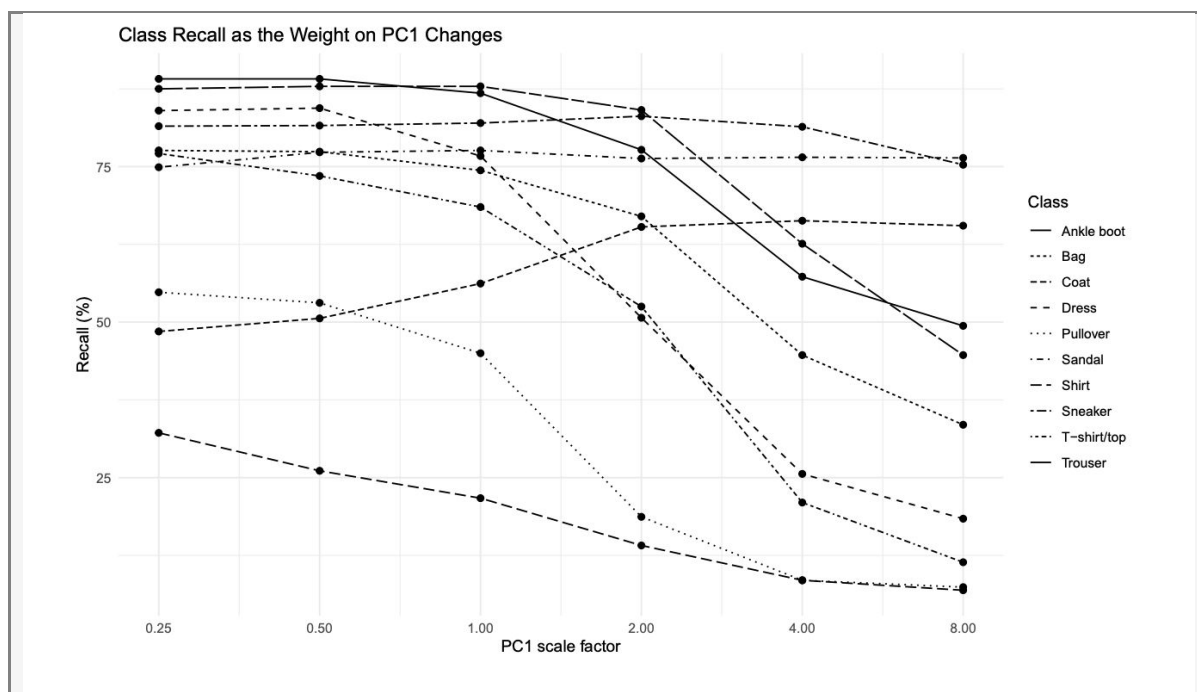


Figure 7. Class-specific recall as the weighting applied to the first principal component (PC1) changes.

3.4 Compression changes computational cost, but projection is not free

There are 7,840 coordinate-distance terms for each test image for the raw classifier. This falls to 840 at 84 dimensions, 240 at 24, and 100 at 10. The comparison burden for downstream comparisons is thus reduced from 78.4 million to 8.4 million, 2.4 million, and 1.0 million test observations, respectively, across 10,000 test observations. The analysis did



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

not, however, include the 784k cost of first projecting the raw vector to a k-dimensional PCA space to avoid the exaggeration of the benefit.

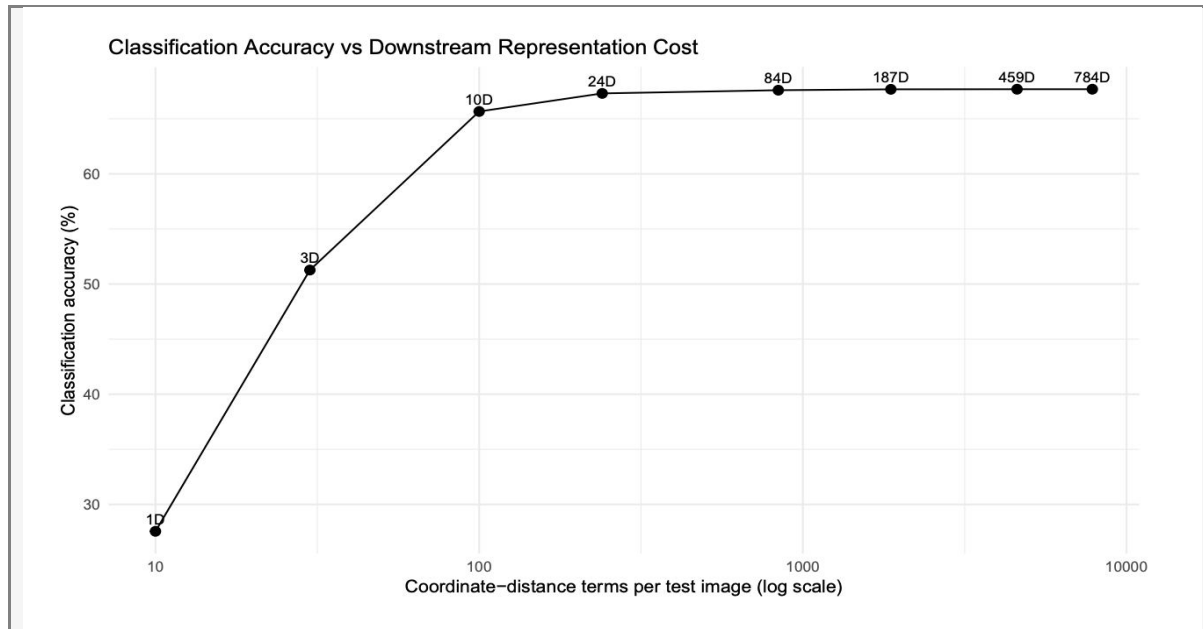


Figure 8. Accuracy as compared to the computation of coordinate-distance downstream.

In the empirical timing exercise, the classification of encoded space was separated from end to end processing. This allowed very small representations to approach the timer resolution, so E2.1 was able to repeat classification operations within larger timing blocks. The resulting precision audit reported median downstream runtimes of 61.02 ms at 784D, 5.64 ms at 84D, 1.94 ms at 24D, 1.20 ms at 10D, and 0.85 ms at 3D, corresponding to speedups of 10.82 $\times$, 31.45 $\times$, 50.85 $\times$, and 71.79 $\times$ relative to 784D.

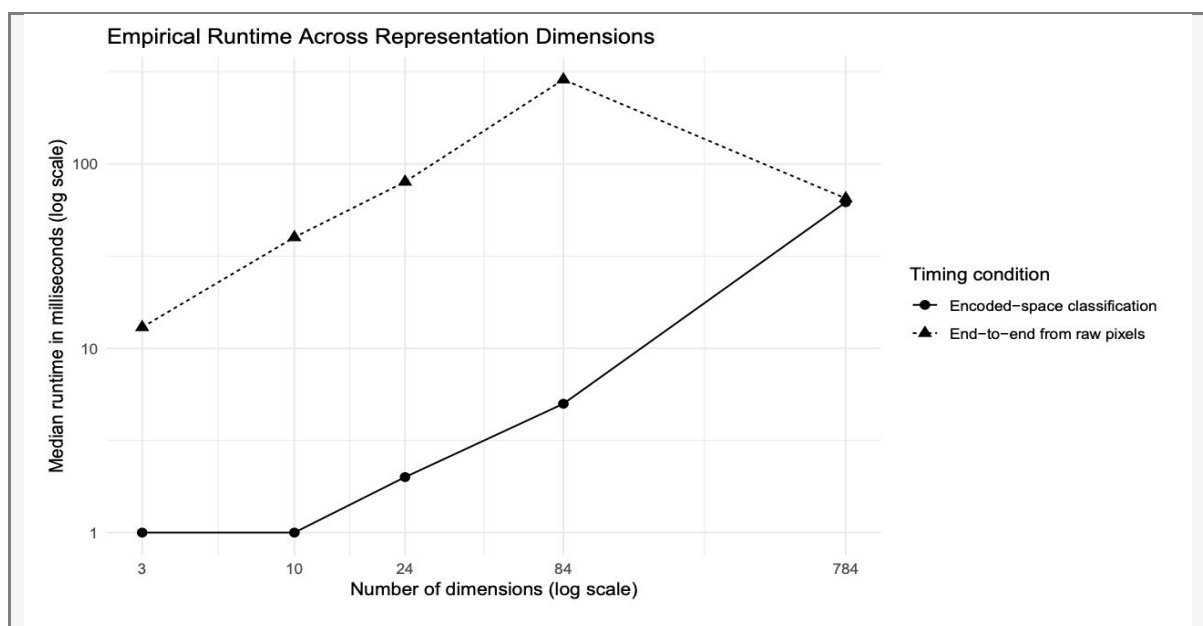


Figure 9. Empirical runtime on representation dimensions: encoded space vs end to end computation from raw pixels.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The reuse analysis then questioned how long it will take to pay back the one time cost of the PCA projection. The 84-dimensional representation broke even after about 5.020 uses, or 6 full uses, requiring the cost of six additional uses; 24D required an additional use, or about 1.490, to pay for itself; 10D and 3D would pay for themselves within an additional use.

IV. DISCUSSION

Combined, these experiments demonstrate the need to view linear algebra as a component of algorithmic thinking only if one does not simplify it to mathematics as a prerequisite. The first of the two is representational, the visual object can be represented as a vector, and the relationships between the vectors have substantial semantic structure in them. The second is geometric: what is considered “near” will change based on the metric, and that will affect real predictions. The third is spectral: The matrix may have full algebraic rank, but it contains most of the useful information in comparatively few directions. Most important of all is the separation of information from geometry in the transformation experiments. The meaning of all the numerical coordinates have changed, but Euclidean relationships and all predictions remain the same with an orthogonal rotation. In contrast, direction-selective scaling is invertible and does not erase information per se, but produces significant changes in distances and classifier decisions. The algorithm is then not just reacting to the information content, but to the way that the information is represented that makes it geometrically significant. The PCA experiments also separate out the variance due to the task from the variance due to the total variance. Only 80.11% of the total variance in the training space is retained by 24 components but 99.44% of the full space classification accuracy is retained. The other variability is not equally important to this specific decision rule. This is in line with the general idea that PCA minimizes the variance in reconstruction, not supervised class discrimination (Jolliffe & Cadima, 2016). The study has some limitations. Fashion-MNIST is under the control and is comparatively small when compared to modern vision tasks. It was deliberately decided to use nearest-centroid classification to be mathematically transparent, rather than to achieve the highest possible accuracy. PCA is linear; more complicated nonlinear representations might reveal other structures. 70.72% result should highlight that the PC1 reweighting experiment is exploratory, as multiple scale factors were evaluated on the same test set and it should not be read as an independently validated optimized classifier. Last but not least, runtime results will vary from one hardware/implementation to another. The main use of their principal value is not to provide universal millisecond benchmarks, but rather to provide a relative computational pattern.

V. CONCLUSION

The analysis shows a smooth continuum between linear algebra and algorithmic behaviour. The 784-dimensional vectors of the Fashion-MNIST images demonstrated class similarity and classification error via geometric vector analysis; the eigendecomposition of a formally full-rank matrix showed that the majority of the variance was concentrated in a small number of dimensions; PCA removed over 96% of the dimensions with very little loss of nearest-centroid accuracy; the same algorithm applied to orthogonal transformations maintained its behavior exactly; selective geometric re-weighting changed the results of thousands of decisions; and reduced representations significantly decreased the cost of computation down the stream. Algorithms are not going to deal with abstract objects, they are going to deal with representations, so linear algebra is therefore essential to algorithmic thinking. Vectors are used to encode observations, metrics are used to define similarity, matrices are used to reveal structure, transformations are used to define the types of relationships preserved, and dimensionality is related to the information in an observation and also determines the computational burden. An understanding of these operations can allow us not just to run an algorithm, but to also explain its behavior.

REFERENCES

1. Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). *On the surprising behavior of distance metrics in high dimensional spaces*. In Database Theory - ICDT 2001 (Lecture Notes in Computer Science, Vol. 1973, pp. 420–434). Springer. https://doi.org/10.1007/3-540-44503-X_27.
2. Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for Machine Learning*. Cambridge University Press. <https://doi.org/10.1017/9781108679930>.
3. Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211–218. <https://doi.org/10.1007/BF02288367>.
4. Golub, G. H., & Van Loan, C. F. (2013). *Matrix Computations* (4th ed.). Johns Hopkins University Press.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

5. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, 498–520. <https://doi.org/10.1037/h0071325>.
6. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>.
7. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572. <https://doi.org/10.1080/14786440109462720>.
8. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. <https://doi.org/10.1145/361219.361220>.
9. Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*. arXiv:1708.07747



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details