Revisiting Dimensionality Reduction with Autoencoders: A Modern Empirical Replication

Wilton Miller
University of Toronto
January 10th, 2026

ABSTRACT
We revisit the comparison between principal component analysis and nonlinear autoencoders for dimensionality reduction, originally studied by Hinton and Salakhutdinov in 2006. Using the MNIST dataset and a modern training setup without layer-by-layer pretraining, we evaluate reconstruction performance across different latent dimensions. Autoencoders are trained using gradient-based optimization and compared against PCA using a shared reconstruction error metric. It is shown that nonlinear autoencoders achieve a lower reconstruction error than PCA for moderate latent dimension sizes, supporting the claims of the original work under contemporary optimization methods. This reading is intended as a modern empirical replication, rather than an exact historical reproduction.

1 INTRODUCTION

Dimensionality reduction is a core problem in machine learning; it has applications in compression, visualization, and representation learning. Simplistic and common approaches, such as principal component analysis (PCA), provide linear projections that preserve variance in the data and give closed-form solutions, making them widely used and well understood.

Autoencoders provide a nonlinear alternative to PCA by learning an encoder and decoder pair. This pair maps input data into a low-dimensional *latent space* and then reconstructs it back to the input space. When restricted to linear transformations, autoencoders are equivalent to PCA. Using non-linear activation functions, however, allows the autoencoder to capture structure that cannot be represented by linear projections.

In their 2006 paper, Hinton and Salakhutdinov (Hinton and Salakhutdinov, 2006) showed that deep nonlinear autoencoders could substantially outperform PCA in reconstruction error across several datasets, including MNIST. An important contribution of their work was showing that effective optimization - by treating each layer as a Restricted Boltzmann Machine (RBM) - was essential for achieving this advantage.

Since 2006, advances in optimization algorithms, activation functions, and compute, have significantly changed the practical training landscape for neural networks. This begs the question of whether the perceived advantages of autoencoders over PCA remain under training methods that do not rely on specialized pretraining procedures.

In this work, we conduct a modern empirical replication of the PCA versus autoencoder comparison on MNIST, focusing on reconstruction error across a range of latent dimensions. Our goal is not to reproduce the original training pipeline, but to decide whether its central empirical claim holds under contemporary modeling choices.

2 METHODS

Dataset

All experiments are conducted on the MNIST handwritten dataset (LeCun et al., 1998), consisting of 60,000 training images and

10,000 test images. Each image is flattened into a vector $x \in \mathbb{R}^{784}$, and pixel intensities are scaled to fall in $[0, 1]$. Reconstruction performance is evaluated on the test set using mean squared error (MSE). Define the dataset to be:

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^{n}, \qquad x^{(i)} \in \mathbb{R}^{D}, \quad D = 784. \tag{1}$$

Reconstruction Objective
Both PCA and autoencoders are evaluated using MSE. For an input x and its reconstruction $\hat{x}$, the expected risk is:

$$\mathcal{L} = \mathbb{E}_{x\sim\mathcal{D}}\left[\|x - \hat{x}\|_2^2\right] . \tag{2}$$

In practice, this expectation is empirically approximated over a dataset of size m by:

$$\widehat{\mathcal{L}} = \frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - \hat{x}^{(i)}\|_2^2. \tag{3}$$

Principal Component Analysis (PCA)
PCA is used as a linear baseline for dimensionality reduction. It assumes the data lies near a linear subspace passing through the origin, as a result it operates on centered inputs. If it were not to operate on centered inputs, the first principal component would largely capture the mean of the data, rather than its variance. Let the empirical mean and centered data be:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}, \qquad \tilde{x}^{(i)} = x^{(i)} - \mu. \tag{4}$$

We then define the covariance matrix as:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}^{(i)}\tilde{x}^{(i)\top} = \frac{1}{n}X_c^\top X_c. \tag{5}$$

where X_c denotes the matrix of centered data points.

PCA then proceeds by computing the spectral decomposition of the covariance matrix:

$$\Sigma = Q\Lambda Q^\top, \tag{6}$$

where the columns of $Q = [u_1, u_2, ..., u_D]$ are the eigenvectors of Σ, and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_D)$ contains the corresponding eigenvalues, ordered such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D$. Then let:

$$U_d = [u_1, u_2, \ldots, u_d] \in \mathbb{R}^{D\times d}. \tag{7}$$

be the matrix of the top d principal components. Now, given an input x, PCA computes the low-dimensional representation:

$$z = U_d^\top (x - \mu), \tag{8}$$

and reconstructs the input as follows:

$$\hat{x} = \mu + U_d z = \mu + U_d U_d^\top (x - \mu). \tag{9}$$

This reconstruction corresponds to the optimal rank-d linear approximation of the data, i.e. minimizing the MSE between the actual input and its reconstruction.

Autoencoder Model
Autoencoders are nonlinear encoder → decoder models that learn to map inputs to a low-dimensional latent representation and reconstruct them back to the input space. Formally, an autoencoder consists of an encoder:

$$f_\theta : \mathbb{R}^D \to \mathbb{R}^d \tag{10}$$

and a decoder:

$$g_\phi : \mathbb{R}^d \to \mathbb{R}^D \tag{11}$$

Given and input x, the latent representation and reconstruction are defined as follows:

$$z = f_\theta(x), \qquad \hat{x} = g_\phi(z) = g_\phi(f_\theta(x)). \tag{12}$$

The parameters θ and ϕ are learned by minimizing the empirical reconstruction loss:

$$(\theta^*, \phi^*) \in \arg\min_{\theta,\phi} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - g_\phi\big(f_\theta(x^{(i)})\big) \right\|_2^2 . \tag{13}$$

In our experiments, both the encoder and decoder are multilayer perceptrons with ReLU activations in the hidden layers and sigmoid activation in the output layer to match the $[0, 1]$ pixel range. All autoencoder models are trained end-to-end using the Adam optimizer (Kingma and Ba, 2015) to minimize the reconstruction loss.

Relation Between PCA and Autoencoders
When both the encoder and decoder are restricted to linear maps and the inputs are centered, the autoencoder reconstruction objective reduces to PCA. In this case, the optimal solution spans the same principal subspace obtained from the spectral decomposition of the covariance matrix, up to a change of basis in the latent space. Introducing nonlinear activations allows autoencoders to represent structure not captured in the linear subspaces defined by PCA.

3 EXPERIMENTS

Experiment Protocol
All experiments are conducted on the MNIST dataset using a common experimental setup across methods. PCA and autoencoders are evaluated at the same latent dimensions $d \in \{2, 16, 32, 64\}$.

For PCA, a separate model is fit for each latent dimension using the training set only. Reconstructions are generated using the inverse transform and are evaluated on the test set.

For autoencoders, a fixed MLP architecture is used across all experiments, with only the bottleneck dimension changed. Each model is trained end-to-end using gradient-based optimization for a fixed number of epochs. No layer-wise pretraining was done, as well as no regularization techniques, or architectural modifications applied across latent dimensions.

All autoencoder runs use a single random seed and a fixed batch size. Training and evaluation are performed using identical data reprocessing and evaluation code. Reconstruction performance is evaluated exclusively on the MNIST test set.

Evaluation Artifacts
Quantitative performance is measured using MSE reconstruction error on the test set for each method and latent dimension. These results are shown in a single plot where test MSE is a function of latent dimensionality.

Qualitative reconstruction grids are generated for a fixed, deterministic subset of test images. These grids are used to visually compare reconstructions produced by PCA and autoencoder at the specific latent dimensions.

4 RESULTS

Figure 1 shows the mean squared reconstruction error on the MNIST test set as a function of the latent dimensionality. The results are for the latent dimensions $d \in \{2, 16, 32, 64\}$.

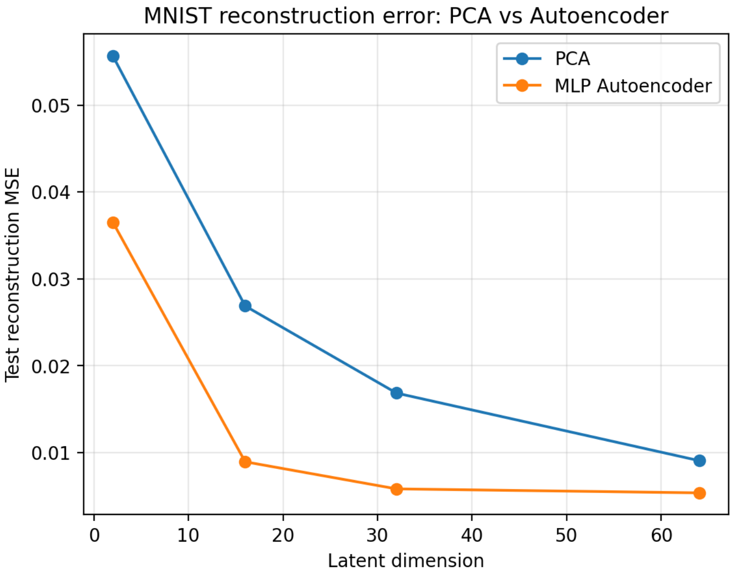
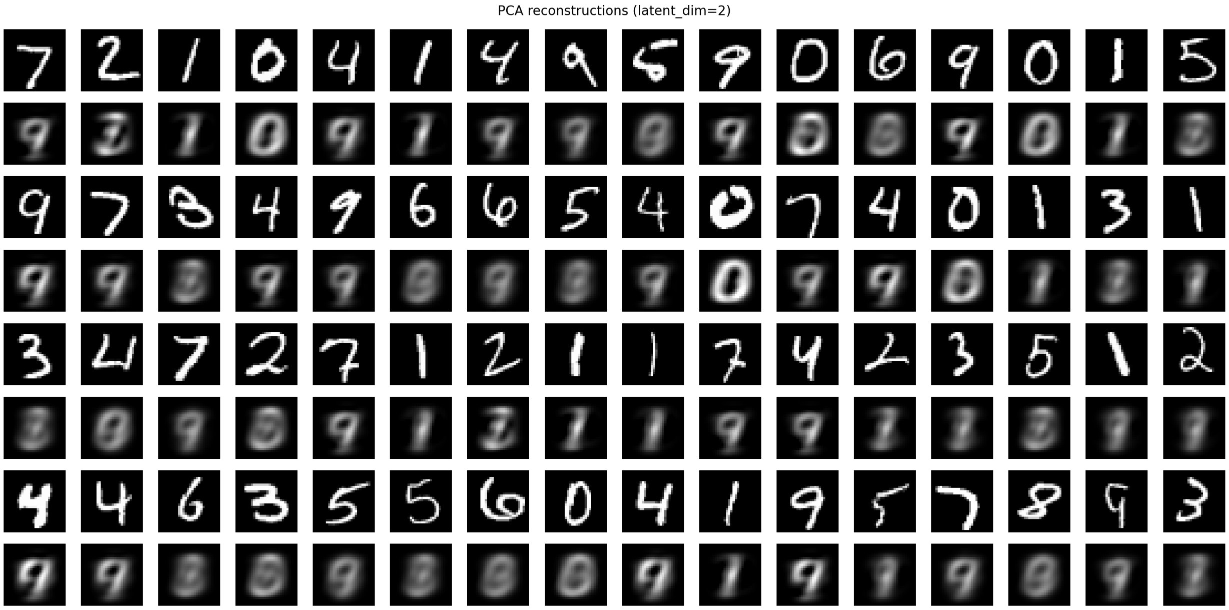


Figure 1: Test-set mean squared reconstruction error as a function of latent dimension for PCA and a nonlinear autoencoder on MNIST. Lower values indicate better reconstruction.

For both methods, reconstruction error decreases monotonically as the latent dimensionality increases. At the smallest latent dimension ($d = 2$), PCA and the autoencoder have largely different reconstruction errors. As the latent dimensionality increases, the autoencoder achieves a lower test MSE than PCA, with the largest absolute different observed at intermediate dimensions. At higher latent dimensionality ($d = 64$), reconstruction error for both methods is reduced, and the gap between PCA and the autoencoder is noticeable more narrow.

The qualitative reconstruction grids are consistent with the quantitative results. At low latent dimensionality, both methods produce reconstructions that capture only coarse digit structure. At higher dimensions, the autoencoders reconstructions preserve shaper digit contours and diner visual details compared to PCA.



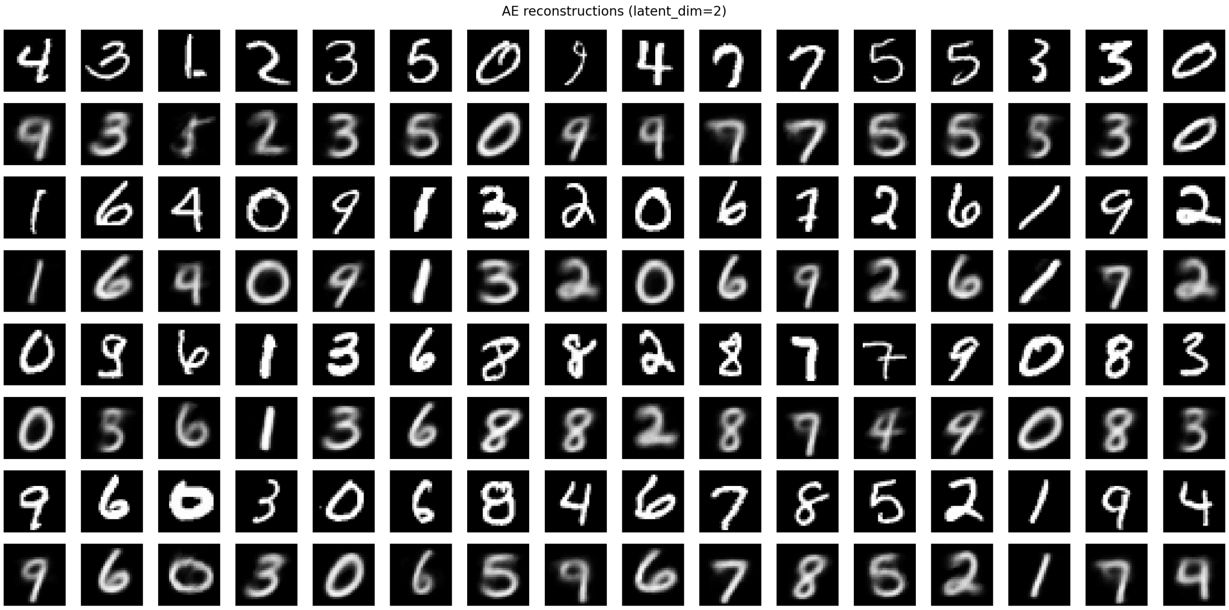
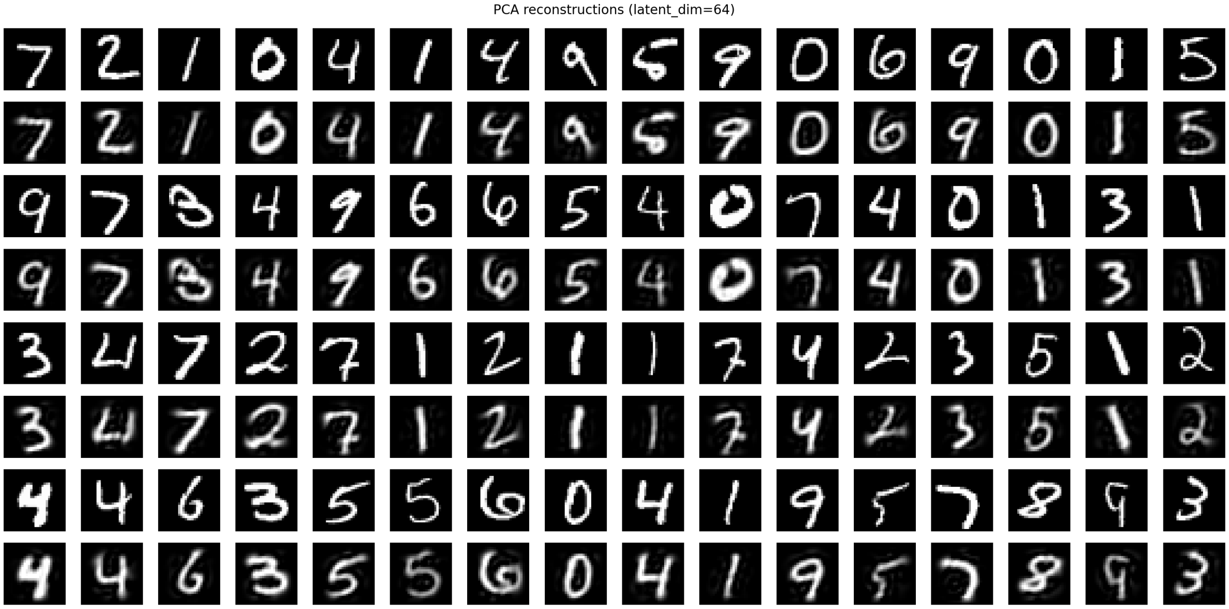


Figure 2: Qualitative reconstructions on MNIST at latent dimension $d = 2$. For each figure, the rows alternate real → reconstruction. The top figure shows the reconstructions produced by by PCA, and the bottom figure shows the reconstructions produced by the autoencoder. Reconstructions are shown for the same fixed subset of test images.



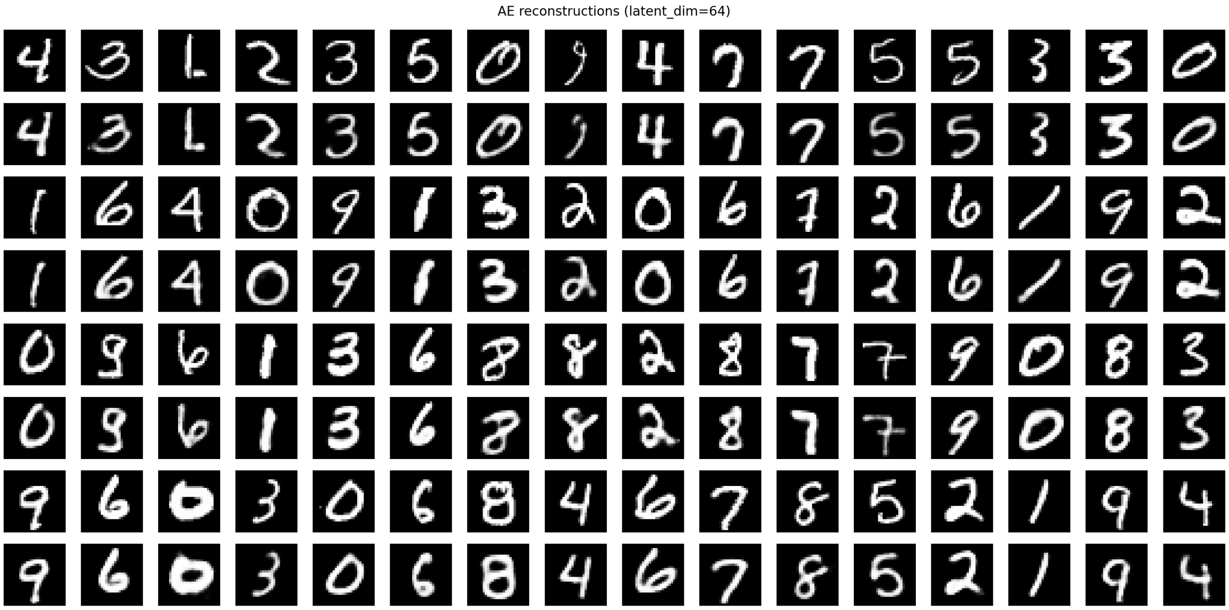


Figure 3: Qualitative reconstructions on MNIST at latent dimension $d = 64$. For each figure, the rows alternate real → reconstruction. The top figure shows the reconstructions produced by by PCA, and the bottom figure shows the reconstructions produced by the autoencoder. Reconstructions are shown for the same fixed subset of test images.

5 DISCUSSION

The results presented in this study reproduce the main empirical observation made by Hinton and Salakhutdinov in 2006: deep nonlinear autoencoders can achieve a lower reconstruction error than PCA when evaluated at the same latent dimensions. Importantly, this claim is observed under a modern training setup that does not require layer-wise pretraining.

A key difference from the original work is the optimization procedure. Whereas Hinton and Salakhutdinov used Restricted Boltzmann Machines to initialize the autoencoders, the models considered here are trained end-to-end using modern gradient-based optimization. The persistence of the qualitative reconstruction advantage under these conditions suggests that the phenomenon is not specific to the historical training pipeline, but continues to be observable under current modeling practices.

This study is intentionally limited in scope. Experiments are conducted on a single dataset, using a single model architecture from a single random seed. Latent dimensionality is varied, but other hyperparameters are held fixed, and no attempt is made to tune models for optimal performance at each dimensionality, as reconstruction error is the sole evaluation metric.

As such, the results should be interpreted as a limited empirical replication rather than a comprehensive comparison of dimensionality reduction methods. The goal is not to establish a single best model for the task in question, but to assess whether a key qualitative result from the original study remains observable under contemporary training conditions.

6 CONCLUSION

This work presented a modern replication of the comparison between PCA and nonlinear autoencoders originally studied by Hinton and Salakhutdinov in 2006. Using MNIST and contemporary training methods without layer-wise pretraining, reconstruction performance was evaluated across a range of latent dimensionalities.

The results show that nonlinear autoencoders achieve lower reconstruction error than PCA at matched latent dimensions under these conditions, consistent with the qualitative findings of the original study.

REFERENCES

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2006.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.