

Empirical Verification of the Variance–Reconstruction Error Equivalence in PCA

ECE 57000 AI Course Project – TinyReproductions Track • Amin Behnami

Motivation and Problem

- High-dimensional data (images, text, tabular features) are hard to visualize and reason about directly.
- Linear dimensionality reduction methods project data into a lower-dimensional subspace.
- Principal Component Analysis (PCA) is a classic method that chooses directions of maximum variance.
- A theoretical result says: for mean-centered data and fixed dimension k , PCA both
 - maximizes captured variance, and
 - minimizes squared reconstruction error among all rank- k linear projections.

Experimental Setup

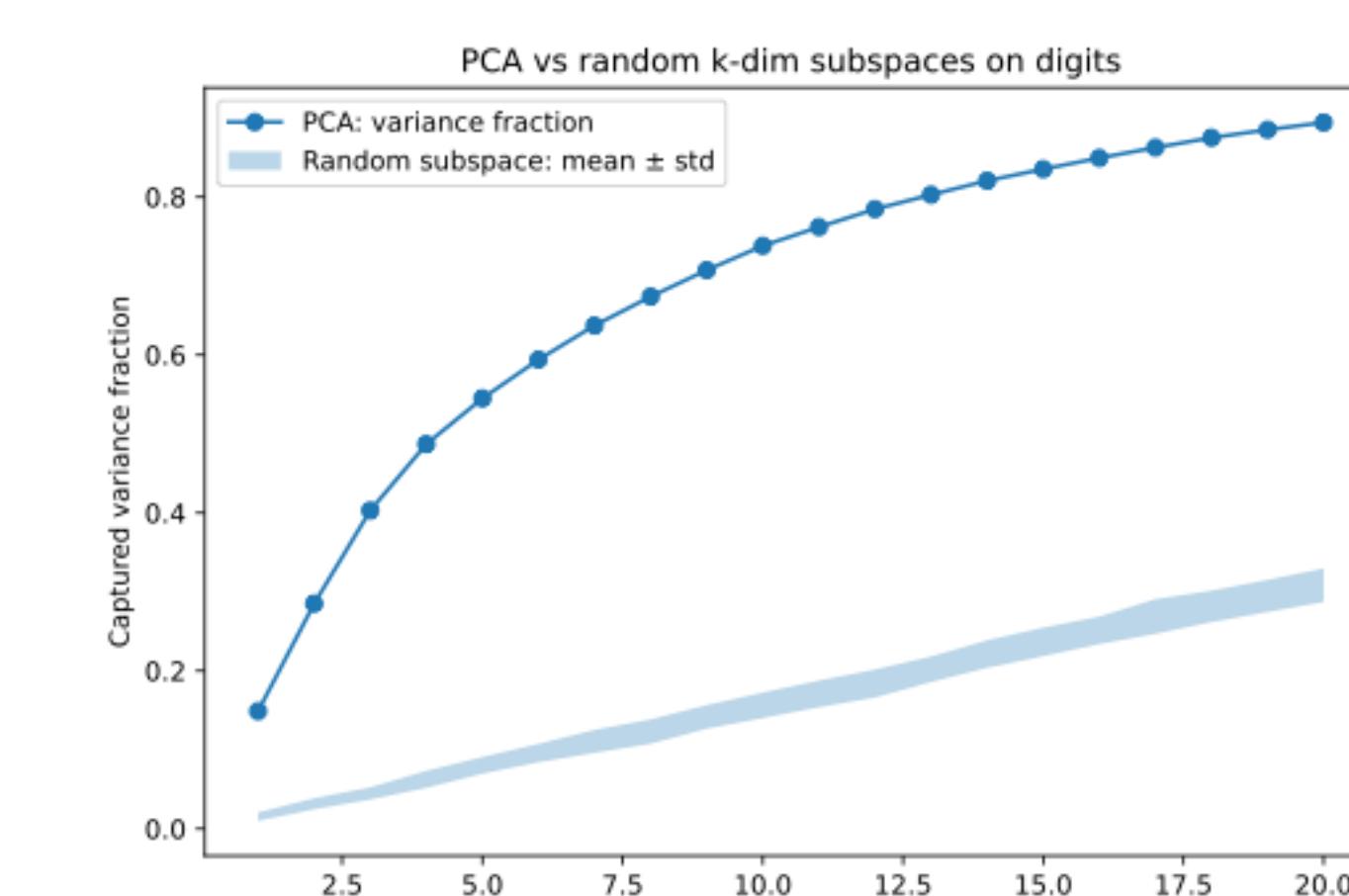
- Dataset: scikit-learn digits ($n \approx 1797$, $d = 64$).
- Preprocessing: mean-center each feature to obtain X_c .
- For each $k \in \{1, \dots, 20\}$:
 - PCA: fit PCA and keep top k components.
 - Random subspaces: sample Gaussian matrix, orthonormalize columns by QR to get a random k -dimensional subspace.
- Metrics:
 - captured variance fraction = variance of projection / total variance
 - reconstruction quality = $1 - (\text{mean-squared reconstruction error} / \text{total baseline MSE})$

Reference:
Kevin P. Murphy, "Machine Learning: A Probabilistic Perspective," MIT Press, 2012.

On the digits dataset,
PCA captures more
variance and achieves
lower reconstruction
error than random
linear subspaces for
every k

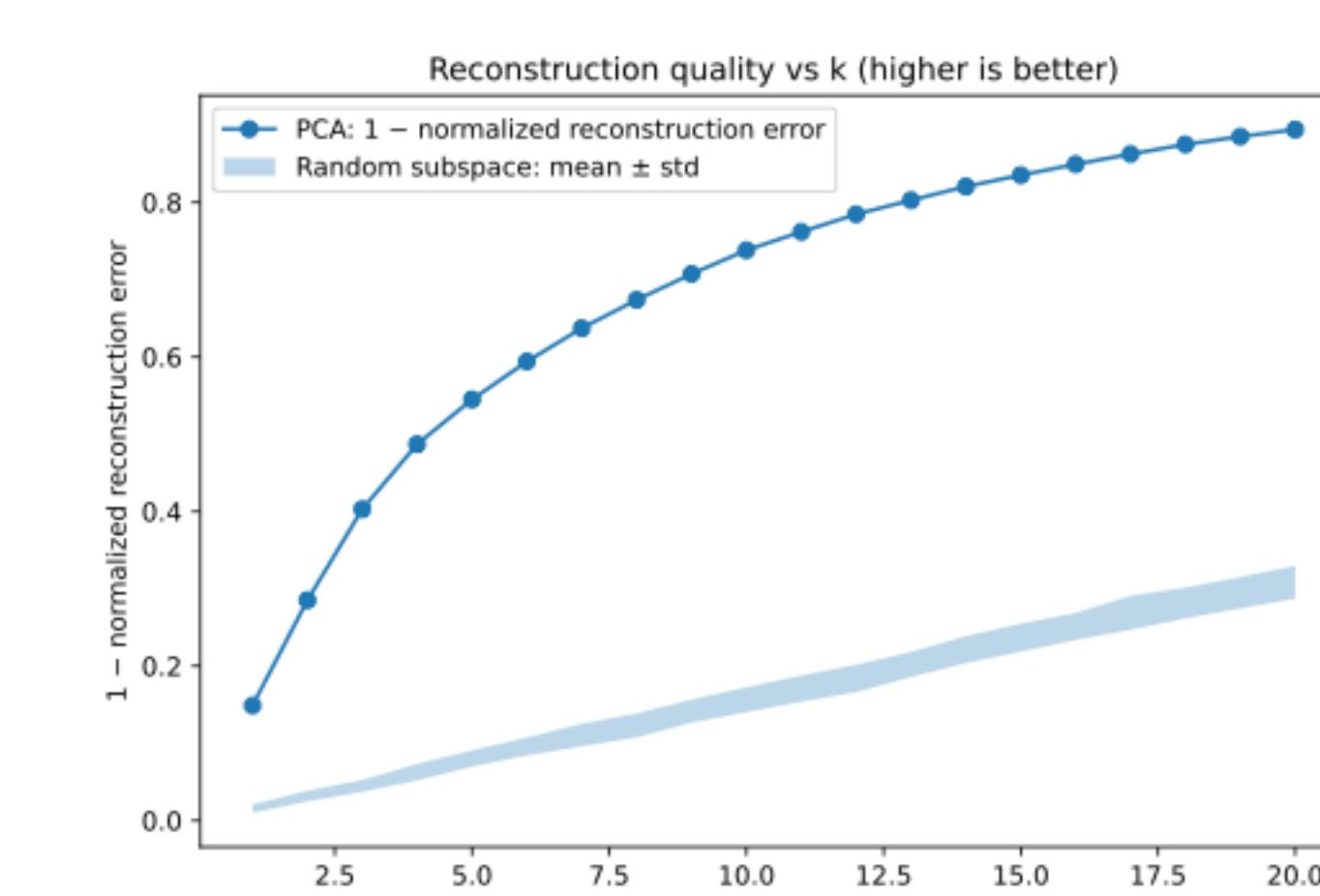
Results: Variance vs k

- Fraction of total variance captured vs k for PCA and random orthonormal k -dimensional subspaces.
- PCA curve lies above the random-subspace mean for all $k = 1, \dots, 20$.
- The gap is largest for small k , where the choice of subspace matters most.
- As k increases, both PCA and random subspaces approach full variance, so curves converge.



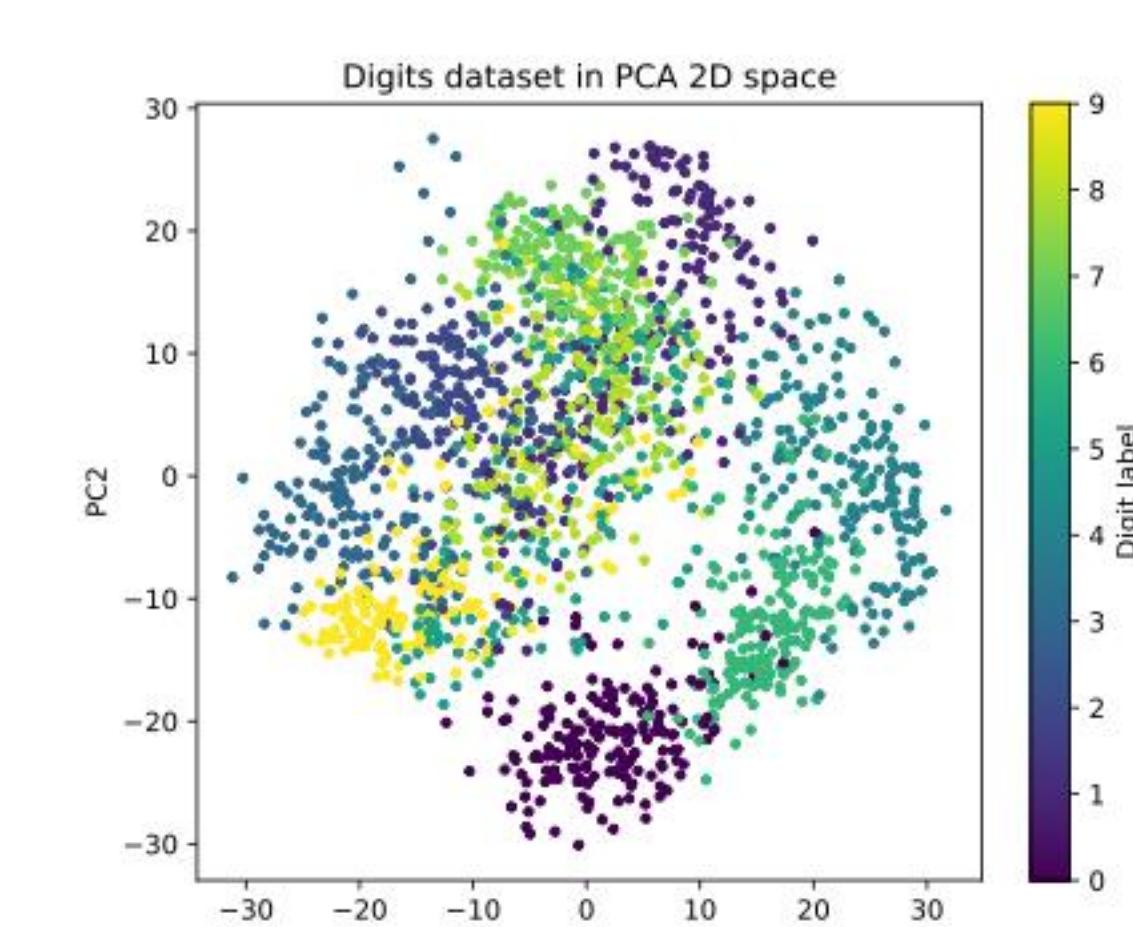
Results: Reconstruction Quality vs k

- Metric: reconstruction quality = $1 - \text{normalized mean-squared error}$.
- PCA dominates the random baseline here as well for every k .
- Variance and reconstruction are complementary: capturing more variance means lower reconstruction error.



Visualization: 2D PCA Embedding

- Projection onto the first two principal components reveals clusters by digit label.
- Even in 2D, several digits form distinct regions, showing PCA finds meaningful structure.



Conclusion and Limitations

- This tiny reproduction empirically confirms that PCA is extremal for both variance and reconstruction error on the digits dataset.
- Limitations: single dataset ($d = 64$), only linear PCA, random isotropic baselines, in-sample evaluation.
- Future work: apply to other datasets, add train/test splits, compare with truncated SVD and nonlinear methods.