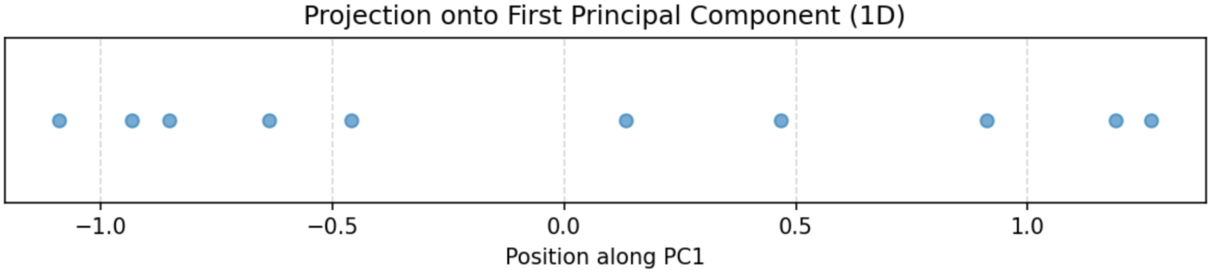
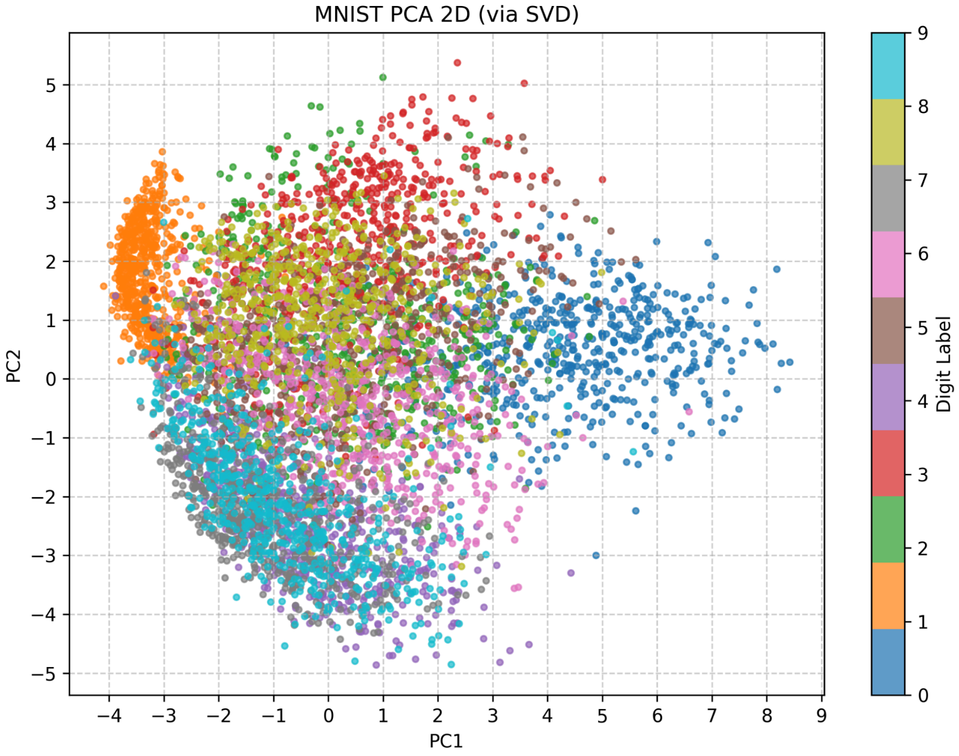
BME 205 Assignment 5 Explanation

1D Projection



This numberline plot shows the distribution of our small sample data along the first principal component, or the eigen vector with largest variance/associated eigen value. The principal components calculated through sklearn and those derived from our covariance matrix were the same ([0.42, 0.91]), with identical lambda ratios indicating high variance, as well as equivalent sklearn projection values. Moreover, this vector V1 is noticeably pointing in the same direction of our largest point distribution in our sample data(what would be a line of best fit).

MNIST 2D Plot

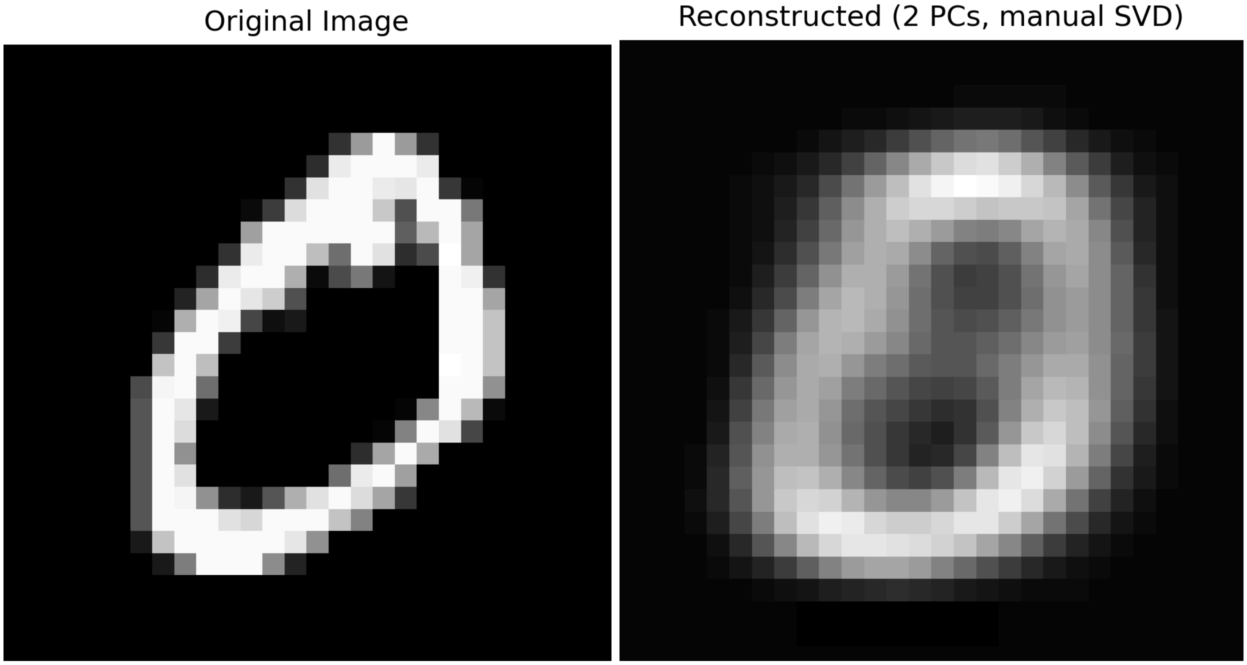


The above image demonstrates the PCA projection along the two largest principal components, or the two eigenvectors with largest associated eigenvalues/variance. In order to get a better grasp of the SVT methods underpinning sklearn’s PCA analysis packages, I decomposed our MNIST sample data matrix into its constituent orthogonal and diagonal matrices.

X = U * Sigma * Vt

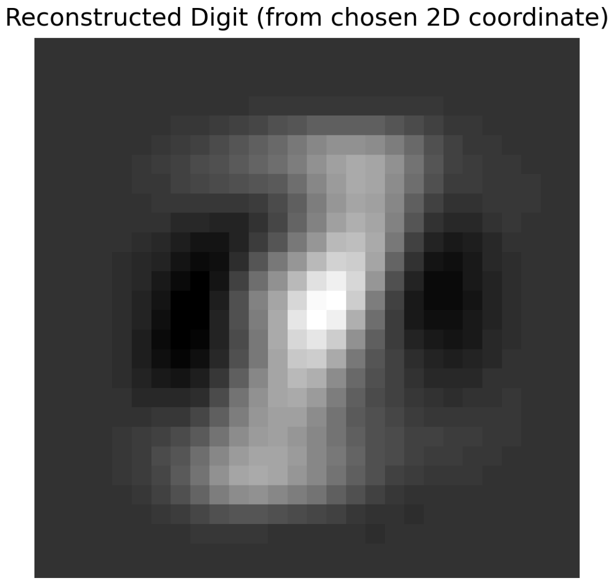
I extracted these three submatrices and utilized the two largest right singular vectors composing V transpose. After centering our dataset X, I projected it along these two PComps to produce a numPy matrix of shape (6000, 2), any of the 6k samples to these principal components, now capable of being plotted in 2D

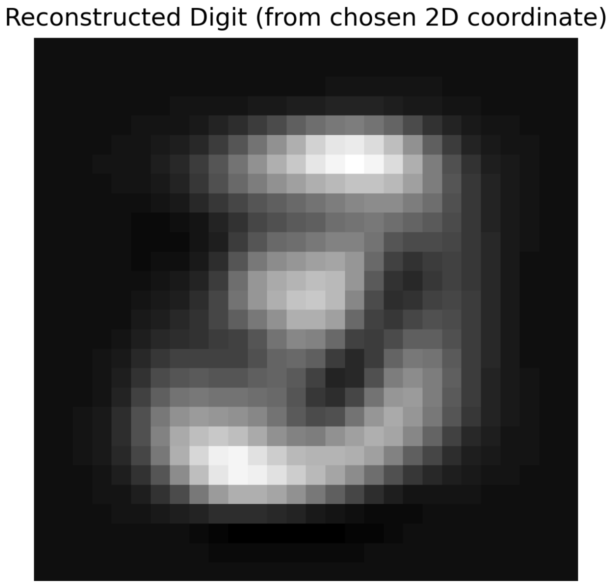
Reconstruction



To determine the efficacy, the first sample of our MNIST dataset was extracted, decomposed with regard to our two Pcomps, and reconstructed via .inverse_transform, which is identical to doing a linear combination of first, the sample’s dot product w.r.t the two PComps and second, the two PComp vectors. By adding the mean back in, the image can then be reconstructed by reshaping to 28 x 28 via .imshow(matplotlib).

The resemblance between the original and the reconstruction is obvious, and this fuzziness in the reconstruction can be attributed to SVD. By only extracting the two largest PComps, there is inherently a disregard for lower variance directions in our feature space. This subset of only two PComps generates a smoother and more generalized reconstruction of our image.


Reconstructed Digit (from chosen 2D coordinate)


Reconstructed Digit (from chosen 2D coordinate)

Along a similar assumption regarding reconstruction, if we were to then manually select a point within our 2D PCA subspace, we can generate numbers that are characteristic to particular digits w.r.t our PComps. On the LHS shows a reconstructed ‘one’ and the RHS shows a ‘three’ selected from the PCA subspace. Again this is not a datapoint but rather a uniquely generated example from this subspace, with fuzziness characteristic to only two PComps.These two examples show strong indication of one and three as they were selected from regions in the projected space highly characteristic to these two values, and these imperfect examples demonstrates the level of accuracy and resolution afforded to subspace of only two PComps.

Dog Ancestry

For our NMF decomposition, our genotype matrix was decomposed into its respective W and H matrices. Wherein W represents ancestry proportions to the n=5 components(normalized and sum to one), and H represents these n=5 ancestry basis mapped back tour 784 genetic SNP feature space. Now averaging ancestry makeup by each clade produced mean ancestry proportions for each respective breed. The .tsv file then lists the n=5 ancestry basis w.r.t each dog clade listed from most significant to least significant.