Report 4: Unsupervised learning: PCA and SOM

Author: *Jannes Nys*

1 Principal Component Analysis on Handwritten Digits

In the Principle Component Analysis, one decomposes the data entries into the principal eigenvectors of the covariance matrix. Each figure is then approximated by a linear combination (for linear PCA) of the limited set of eigenvectors.

Figure 1 shows the total reconstruction error as a function of the number of eigenvectors used in the PCA projection. Figure 1 also shows the total reconstruction error. These figures illustrate that both are proportional. Hence, by setting the required maximum reconstruction error, one can choose a proper number of eigenvectors of the covariance matrix to span the compressed space based on the cumulative sum of the eigenvalues.

When k is equal to 256, which is the dimension of the original basis, the eigenbasis of the PCA spans the full space of the original figures. Hence, the reconstruction error is approximately[1] zero, since the PCA is now a rotation of the figures, which is inverted exactly upon reconstruction.

As an example, we show a reconstructed image for a number of dimensions of the PCA basis in Fig. 2. As the basis increases, the reconstructed image converges towards the original image.

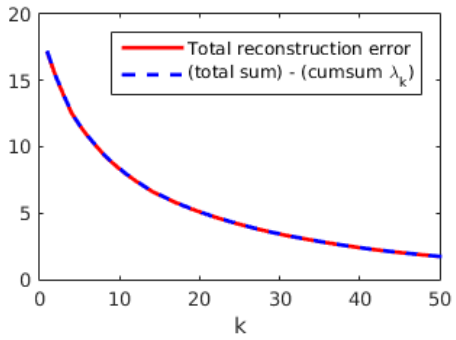

Figure 1: Total sum of all eigenvalues minus the cumulative sum of the reconstruction k largest eigenvalues of the covariance matrix (dashed blue) and mean-squared reconstruction error (solid red) as a function of the dimensionality of the PC basis.

Figure 2: Example of a reconstruction of an image using a basis of (f.l.t.r.) $1, 2, 3$ and 4 principal eigenvectors. The outer-rightmost figure is the original, uncompressed image. It was observed that for this data item, a basis of ~ 50 eigenvectors is required to restore the image properly.

[1] Due to computational round-off errors, during the eigenvector decomposition and reconstruction, the obtained is not exactly 0, but rather 6.7150×10^{-29}.