Principal Components Analysis and Handwritten Digits

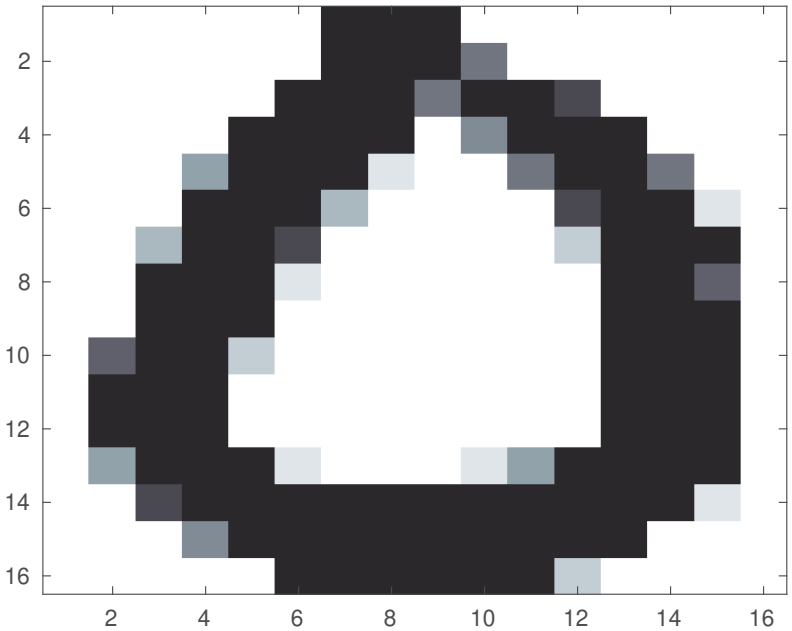
L. Eldén

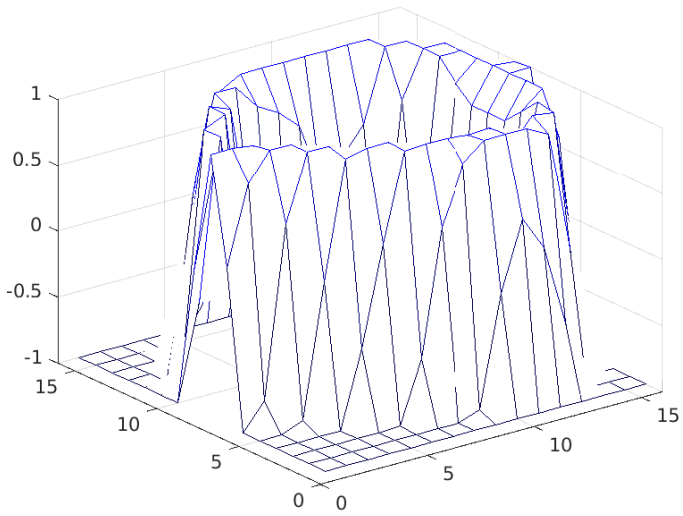
Linköping University

September 2020

Application: Handwritten digits, classification

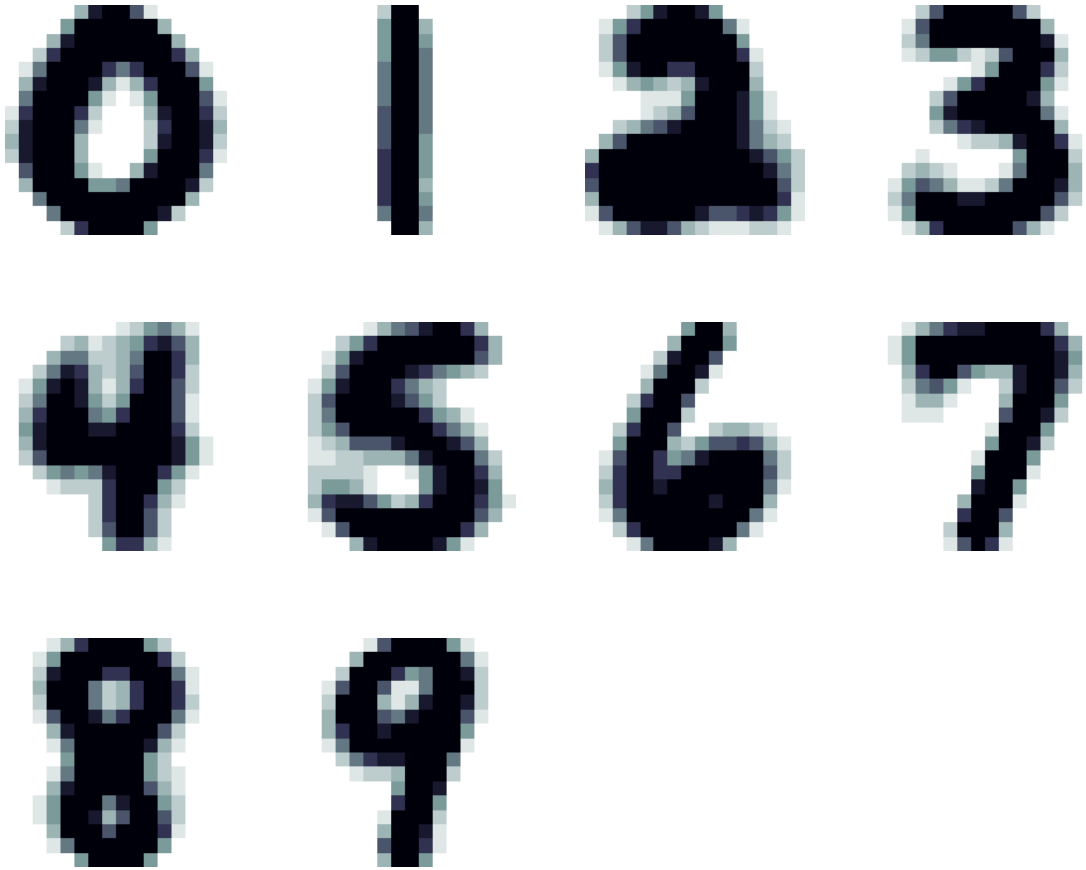
Digitized handwritten digits, 16×16 pixels





Vectorize the image: vector in $\mathbb{R}^{256}$

Means of digits



Set of digits from two classes

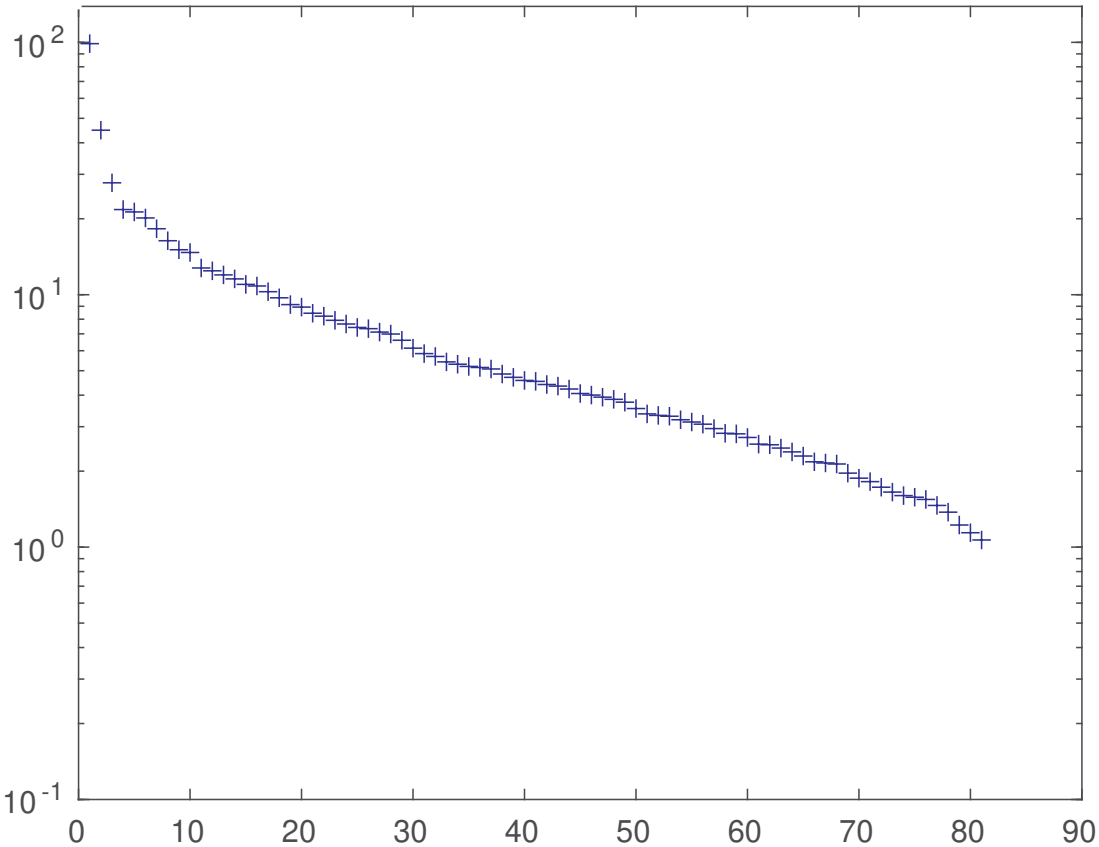
Matrix of digits

81 digits in a matrix $A \in \mathbb{R}^{256 \times 81}$:

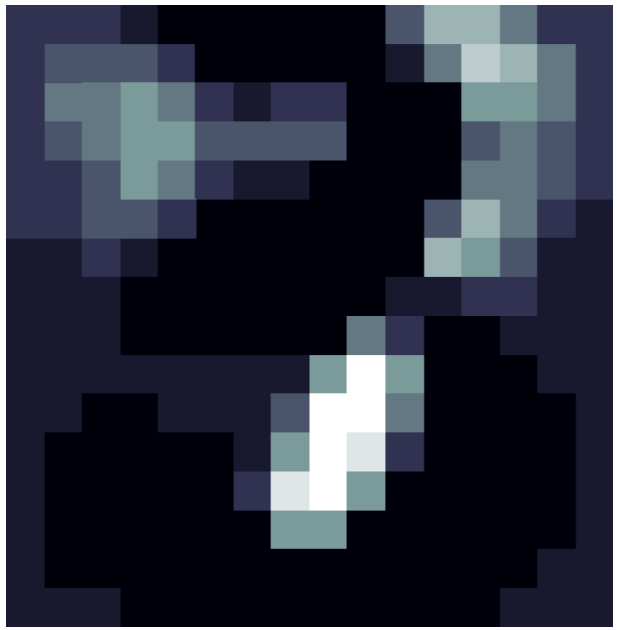
Each column is a digit.

Principal component analysis

Compute the SVD: `[U,S,V] = svd(A);`
Singular values:



First two singular vectors: u_1 and u_2



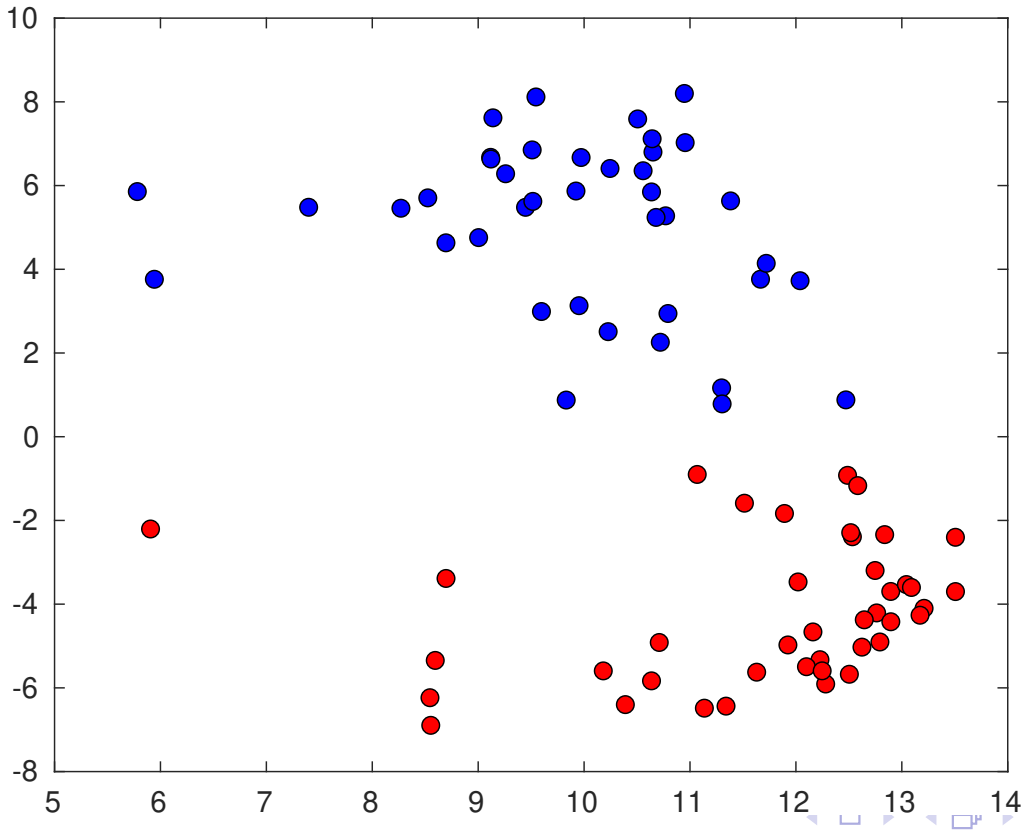
Principal components: u_1 and u_2

PCA

u_1 and u_2 are orthogonal basis vector in a two dimensional subspace of $\mathbb{R}^{256}$

Project all the digits in A down to the 2D space: The coordinates of a vector $a \in \mathbb{R}^{256}$ are $u_1^T a$ and $u_2^T a$

The coordinates of all digits: `(U(:,1:2))'*A`



PCA for classification etc.

Classical approach: Project to a low-dimensional space (2D or 3D) and plot. Separate classes visually.

PCA = SVD analysis

1. compute the parameters of the method (singular vectors) from training data
2. project the data to a k-dimensional space and perform the analysis on test data (that are manually analyzed)
3. select the value of k that gives best performance for the test data
4. apply the method to real data

References

Many different variants are possible, see e.g. the books below.

L. Eldén.
Matrix Methods in Data Mining and Pattern Recognition, Second Edition.
SIAM, 2019.

T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning. Data mining, Inference and Prediction.
Springer, New York, second edition, 2009.
http://www.web.stanford.edu/~hastie/ElemStatLearn/.