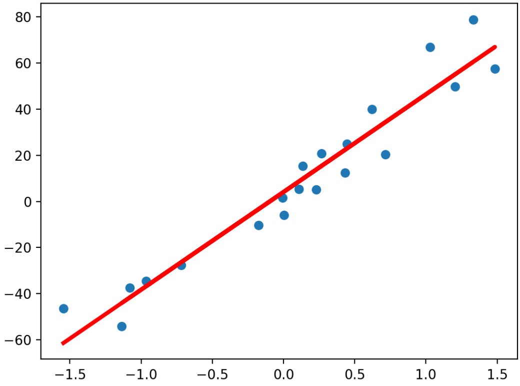
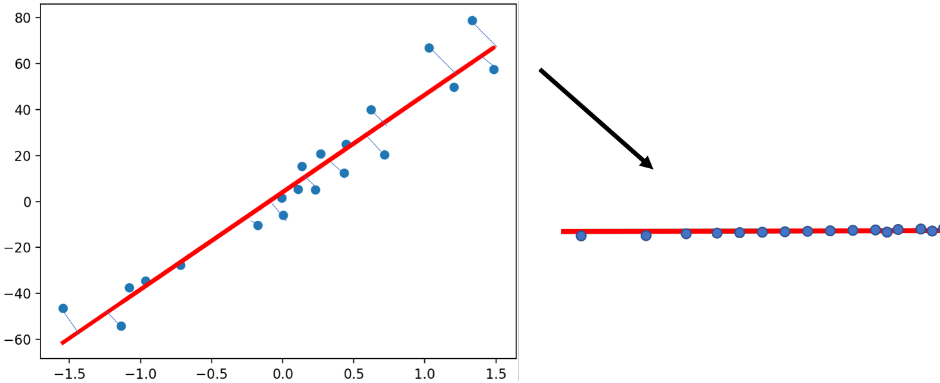
PCA

Sujay Sanghavi

Principal Component Analysis



Principal Component Analysis

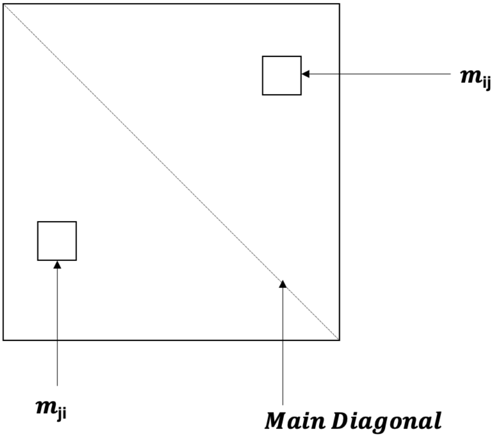


This data in 2D can be represented in 1D without much loss of information.

PCA: Data in higher dimensions represented in lower dimensions while preserving important information.

Matrix Algebra Basics

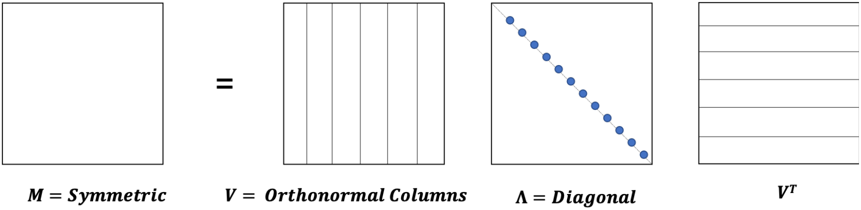
M is a symmetric matrix if it is square and $m_{ij} = m_{ji}$ for all i, j.
i.e. It is equal to its transpose. $M = M^T$.



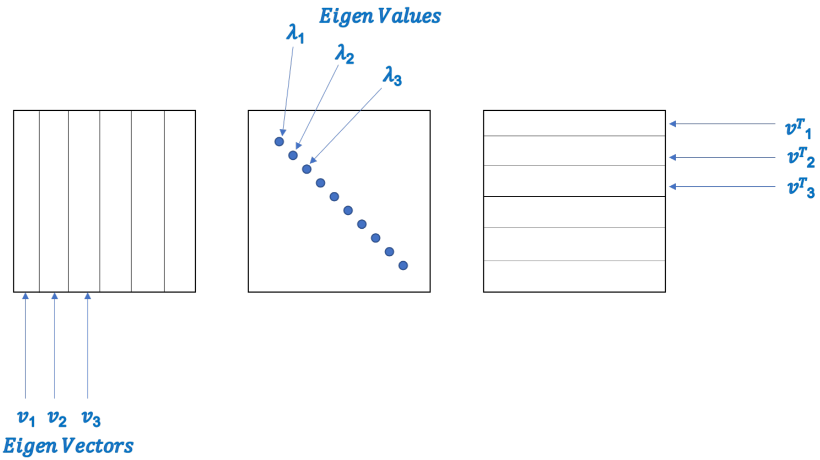
Eigen Values and Eigen Vectors

- Every symmetric matrix can be written as a product:

$$M = V \Lambda V^T$$

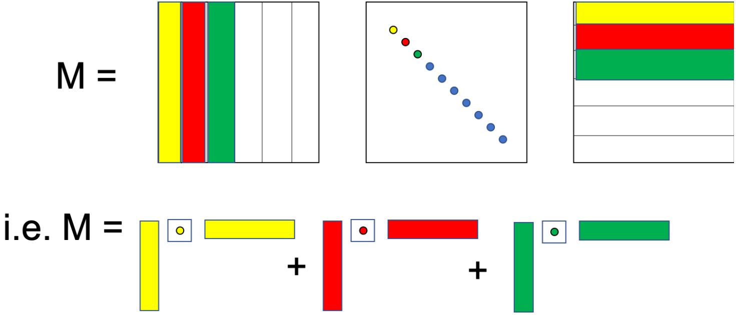


Eigen Values and Eigen Vectors



The eigen vectors are orthonormal, i.e. they are normal: $||v_i||_2^2 = v_i^T v_i = 1$ and they are orthogonal: $v_i^T v_j = 0$.

Eigen Values and Eigen Vectors



Eigen Decomposition of $d \times d$ matrix M:

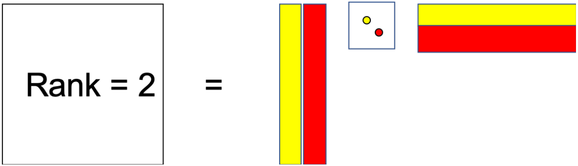
$$M = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \lambda_3 v_3 v_3^T + \dots$$

$$M = \sum_{i=1}^{d} \lambda_i v_i v_i^T$$

Each v_i is a $d \times 1$ vector, each $\lambda_i v_i v_i^T$ is a $d \times d$ matrix.

Rank

- The rank of a symmetric matrix M is the number of non-zero eigenvalues.

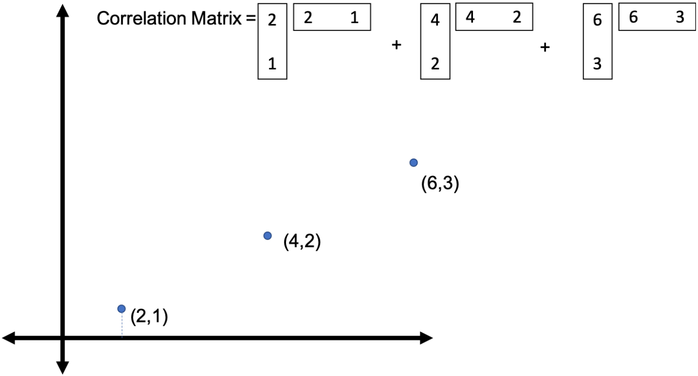


$$M = \sum_{i=1}^{r} \lambda_i v_i v_i^T$$

r: rank

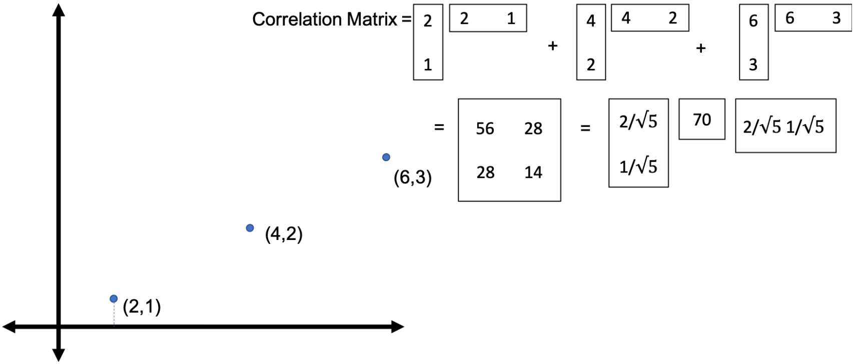
Back to dimensionality reduction

Suppose 3 points lie on a line - "lower dimensional"



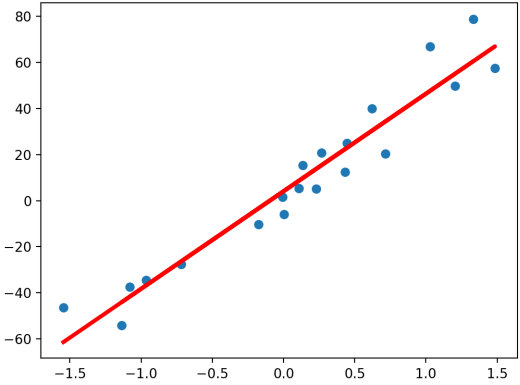
Correlation matrix of the points - write each point as a vector v and sum up vv^T over all points.

Back to dimensionality reduction



Correlation matrix is a symmetric matrix. $(2/\sqrt{5})^2 + (1/\sqrt{5})^2 = 1$. Thus the above is an eigen decomposition with rank 1 as only one eigen vector is required. Note it is rank 1 as the points lie exactly on a line.

Back to dimensionality reduction



Here, correlation matrix will be almost low-rank.

PCA (Symmetric matrix version

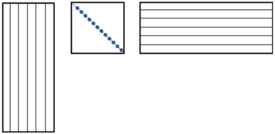
- Approximate a given symmetric matrix by a low-rank matrix

i.e. find rank-r matrix $M^{(r)}$ that minimizes squared error.

$$||M - M^{(r)}||_2^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} |m_{ij} - m_{ij}^{(r)}|^2$$

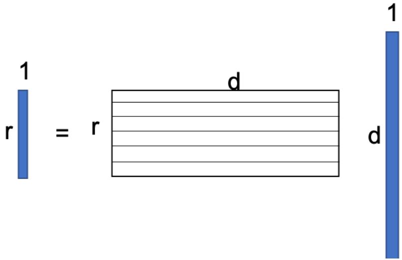
PCA (Symmetric matrix version)

- Given data points $x_1, x_2, \ldots x_n$ and each $x_i \in \mathbb{R}^d$
- Make the correlation matrix $M = \sum_{j=1}^{n} x_i x_i^T$. This is symmetric because
  - Each $x_i x_i^T$ is symmetric $((x_i x_i^T)^T = (x_i^T)^T x_i^T = x_i x_i^T)$
  - sum of symmetric matrices is symmetric
- Take eigen-decomposition of M. $M = \sum_{i=1}^{d} \lambda_i v_i v_i^T = V \Lambda V^T$ where V is $d \times d$
- Reorder eigen values and corresponding eigen vectors such that $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_d$.
- Choose $r < d$ which is the desired rank. We take the top r eigen vectors and discard the others
- Set $M^{(r)} = \sum_{i=1}^{r} \lambda_i v_i v_i^T = V^{(r)} \Lambda^{(r)} (V^{(r)})^T$ where $V^{(r)}$ is a $d \times r$ matrix

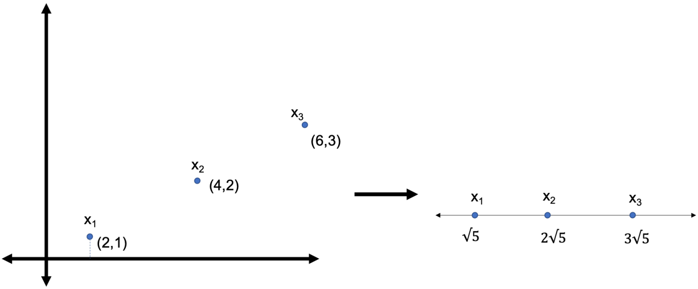


PCA (Symmetric matrix version)

- $\hat{x}_j = (V^{(r)})^T x_j$ i.e. x_j is a $d \times 1$ vector and becomes $\hat{x}_j$ which is an $r \times 1$ vector



Back to dimensionality reduction



$$V = \begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix}$$

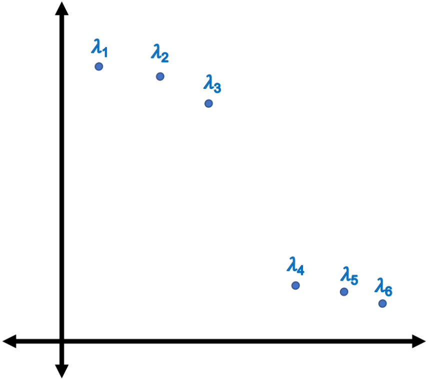
$$\hat{x_1} = \left(2/\sqrt{5} \quad 1/\sqrt{5}\right) \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \sqrt{5}$$

$$\hat{x_2} = \left(2/\sqrt{5} \quad 1/\sqrt{5}\right) \begin{pmatrix} 4 \\ 2 \end{pmatrix} = 2\sqrt{5}$$

If the points were not exactly on a line, then V would have been a 2×2 matrix and we would drop the second vector. Then our final values would be close to but not exactly the values we got above.

Choosing rank

Eigen vectors corresponding to high eigen values are important - carry much of the information about the data. If eigen values become much lower after a point, drop from those eigen vectors onwards.

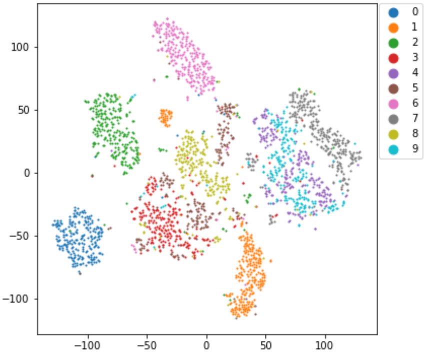


Here we can choose a rank of 3 or more.

Example: MNIST dataset

Correlation matrix of the points - write each point as a vector v and sum up vv^T over all points.

PCA of MNISTt



Correlation matrix of the points - write each point as a vector v and sum up vv^T over all points.