I.9) Principal Components and the Best Low Rank Matrix

The Eckart-Young Theorem states that

$$A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T$$

is the closest rank k matrix to A.

(Here "closest" is with respect to any matrix norm that depends only on the singular values of A, such as the Frobenius norm $\|A\|_F$ or the spectral norm $\|A\|_2$.)

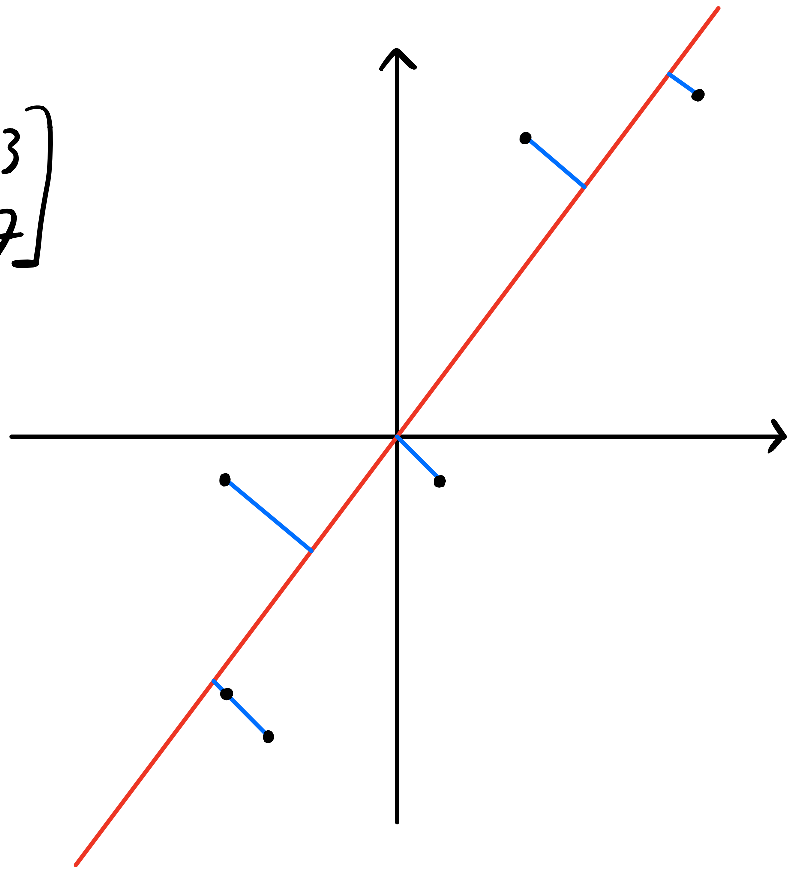
This theorem tells us that k singular vectors explain more of the data than any other set of k vectors: we can choose

$$u_1, \ldots, u_k$$

as the basis for the k-dimensional subspace that is closest to the n data points (ie columns of A).

Example:

$$A = \begin{bmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{bmatrix}$$



A is 2×6:

6 data points in $\mathbb{R}^2$

Note that the points are centered around the origin: their mean is zero. The red line is the span of the singular vector u_1.

This line minimizes the sum of the squared orthogonal distances from the data points to the line.

(This is different than least squares which minimizes the sum of the squared vertical distances from the data points to the line.)

Note that $\{u_1, u_2\}$ are orthonormal. Thus, column j of A can be written as:

$$a_j = (a_j^T u_1) u_1 + (a_j^T u_2) u_2 .$$

Then $\|a_j\|_2^2 = |a_j^T u_1|^2 + |a_j^T u_2|^2$

$$\Rightarrow \sum_{j=1}^{n} \|a_j\|_2^2 = \sum_{j=1}^{n} |a_j^T u_1|^2 + \sum_{j=1}^{n} |a_j^T u_2|^2 .$$

The sum of the squared orthogonal distances is precisely $\sum_{j=1}^{n} |a_j^T u_2|^2$ which we can minimize by maximizing

$$\sum_{j=1}^{n} |a_j^T u_1|^2 = \sum_{j=1}^{n} u_1^T a_j a_j^T u_1 = u_1^T \left(\sum_{j=1}^{n} a_j a_j^T \right) u_1$$

$$= u_1^T (A A^T) u_1 .$$

Which unit vector x maximizes $x^T A A^T x = \|A^T x\|_2^2$? It is the eigenvector of AA^T corresponding to the largest eigenvalue $\lambda_1 = \sigma_1^2$, which is $x = u_1$.

The Statistics Behind PCA

Let $A_0 \in \mathbb{R}^{m \times n}$ be our original data: n points in $\mathbb{R}^m$. For example, if we measure age, height, and weight of 100 people, then $m = 3$ and $n = 100$.

$$
\begin{array}{r c l}
 & \begin{array}{cccc} p_1 & p_2 & \cdots & p_{100} \end{array} & \\
\begin{array}{r} \text{age} \\ \text{height} \\ \text{weight} \end{array} &
\begin{bmatrix} 63 & 41 & & 21 \\ 152 & 157 & \cdots & 156 \\ 48 & 53 & & 54 \end{bmatrix} &
\begin{array}{l} \text{(years)} \\ \text{(cm)} \\ \text{(kg)} \end{array}
\end{array}
$$

The sample mean is $\boxed{\mu = \frac{1}{n}(A_0 e)}$ where $e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

Subtracting the mean from each column of A_0, we get the centered A:

$$\boxed{A = A_0 - \mu e^T}.$$

$$\left[\begin{aligned} \underline{\text{Check:}} \quad Ae &= A_0 e - \mu(e^T e) \\ &= A_0 e - \frac{1}{n}(A_0 e) n \\ &= A_0 e - A_0 e = 0. \quad \checkmark \end{aligned}\right]$$

The <u>covariance matrix</u> of the data
sample

is
$$\boxed{S = \frac{1}{n-1} A A^T}$$
$$= \frac{1}{n-1} \sum_{j=1}^{n} a_j a_j^T$$
$$= \frac{1}{n-1} \sum_{j=1}^{n} (A_{0j} - \mu)(A_{0j} - \mu)^T.$$

The m orthogonal eigenvectors of S are the <u>principal components</u> of A. We compute these principal components by taking the SVD of A.

The <u>total variance</u> is

$$T = \operatorname{trace}(S) = \frac{1}{n-1}\left(\sigma_1^2 + \cdots + \sigma_r^2\right).$$

The proportion of the total variance explained by the first k principal components is $\dfrac{\sigma_1^2 + \cdots + \sigma_k^2}{\sigma_1^2 + \cdots + \sigma_r^2}$ $(k \le r)$.

(numerical demonstration)