University of Edinburgh

INFR11156: Algorithmic Foundations of Data Science (2019)

Lecture 5: Best-fit Subspaces and Singular Value Decomposition (1)

1 Singular values and singular vectors

Let $a = (a_1, \ldots, a_d)$ be a point in $\mathbb{R}^d$. We look at the projection of the points a onto the line through the origin in the direction of v, see Figure 1 for illustration. Then we have that

$$a_1^2 + a_2^2 + \ldots + a_d^2 = (\text{length of projection})^2 + (\text{distance of point to line})^2,$$

and therefore

$$(\text{distance of point to line})^2 = a_1^2 + a_2^2 + \ldots + a_d^2 - (\text{length of projection})^2.$$

Since $\sum_{j=1}^{d} a_j^2$ is constant independent of the line, minimising a's distance to the line is equivalent to maximising its projection onto the line.

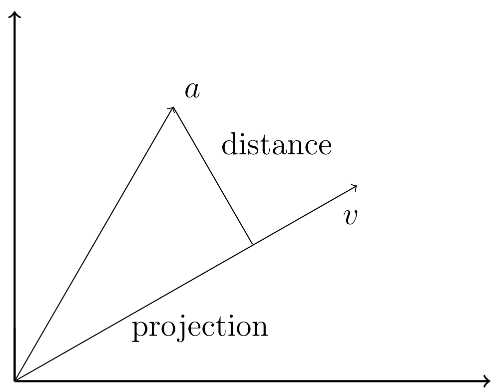

Figure 1: The projection of the point a onto the line through the origin in the direction of v.

Generalising the case above, we assume that there are n points, each of which is represented in $\mathbb{R}^d$. We use matrix $A \in \mathbb{R}^{n \times d}$ to represent these n points. Then, for any fixed direction $v \in \mathbb{R}^d$, the length of the projection of the i-th point, expressed by A_i, is $|\langle A_i, v \rangle| = |A_i v|$, and therefore the best-fit line is the one that maximises $\sum_{i=1}^{n} |A_i v|^2 = \|Av\|^2$. With this in mind, we define the *first singular vector* v_1 of A as

$$v_1 \triangleq \arg\max_{\|v\|=1} \|Av\|.$$

The value $\sigma_1(A) \triangleq \|Av_1\|$ is called *the first singular value* of A. Notice that $\sigma_1^2(A) = \sum_{i=1}^{n} |A_i v_1|^2$ is the sum of the squared lengths of the projections of the points onto the line determined by v_1.

If the data points were all either on a line or close to a line, then intuitively v_1 should give us the direction of that line. However, if the data points are not close to that line but lie close to a 2-dimensional subspace, then further work is needed. We will look at the following greedy approach.

We start by finding v_1 and then find the best 2-dimensional subspace containing v_1. Notice that, for every 2-dimensional subspace containing v_1, the sum of squared lengths of the projections onto the subspace equals the sum of squared projections onto v_1 plus the sum of squared projections along a vector perpendicular to v_1 in the subspace. Hence, instead of looking for the best 2-dimensional subspace containing v_1, we look for a unit vector vector v perpendicular to v_1 that maximises $\|Av\|^2$ among all such unit vectors. This motivates the definition of *the second singular vector* v_2, which is the best-fit line perpendicular to v_1. Formally, we have

$$v_2 \triangleq \arg \max_{v \perp v_1, \|v\|=1} \|Av\|.$$

The value $\sigma_2(A) = \|Av_2\|$ is called *the second singular value of* A. The *third singular vector* v_3 and the *thrid singular value* are defined similarly by

$$v_3 \triangleq \arg \max_{\substack{v \perp v_1, v_2 \\ \|v\|=1}} \|Av\|,$$

and $\sigma_3(A) = \|Av_3\|$.

The greedy algorithm finds the v_1 that maximises $\|Av\|$ and then the best-fit 2-dimensional subspace containing v_1, etc. The following theorem shows that this simple greedy algorithm finds the best-fit subspace of every dimension.

Theorem 1 (The Greedy Algorithm Works)**.** *Let $A \in \mathbb{R}^{n \times d}$ be a matrix with singular vectors $v_1, \ldots, v_r$. For any $1 \leq k \leq r$, let V_k be the subspace spanned by $v_1, \ldots, v_k$. Then, for every k, V_k is the best-fit k-dimensional subspace for A.*

Proof. The proof is by induction. The statement is obviously true for $k = 1$. For $k = 2$, let W be a best-fit 2-dimensional subspace for A. For any orthonormal basis (w_1, w_2) of W, $\|Aw_1\|^2 + \|Aw_2\|^2$ is the sum of squared lengths of the projections of the rows of A onto W. We choose an orthonormal basis (w_1, w_2) of W as follows:

1. If v_1 is perpendicular to W, any unit vector in W that we choose as w_2 is perpendicular to v_1.

2. Otherwise, we choose w_2 to be the unit vector in W perpendicular to the projection of v_1 onto W. This makes w_2 perpendicular to v_1.

Since v_1 maximises $\|Av\|^2$, it holds that $\|Aw_1\|^2 \leq \|Av_1\|^2$. Since v_2 maximises $\|Av\|^2$ overall v perpendicular to v_1, we have that $\|Aw_2\|^2 \leq \|Av_2\|^2$. Thus, we have that

$$\|Aw_1\|^2 + \|Aw_2\|^2 \leq \|Av_1\|^2 + \|Av_2\|^2.$$

Hence, V_2 is at least as good as W and is a best-fit 2-dimensional subspace.

This proof can be used inductively to prove the case for a general k. □

The vectors $v_1, \ldots, v_r$ are called the right-singular vectors. We normalise these vectors and define

$$u_i \triangleq \frac{1}{\sigma_i(A)} A v_i.$$

These u_i are called the left-singular vectors. It is easy to show that u_i similarly maximises $\|u^\intercal A\|$ over all u perpendicular to $u_1, \ldots u_{i-1}$, and these left-singular vectors are also orthogonal.

Lemma 2. *Let $A \in \mathbb{R}^{n \times d}$. Then it holds that $\sum_{i=1}^{n} \sum_{j=1}^{d} a_{ij}^2 = \sum_{i=1}^{r} \sigma_i^2(A)$.*

Proof. Let $v_1, \ldots, v_r$ be the singular vectors of A. Then it holds that

$$\sum_{j=1}^{n} \|A_j\|^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (A_j v_i)^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} (A_j v_i)^2 = \sum_{i=1}^{r} \|Av_i\|^2 = \sum_{i=1}^{r} \sigma_i^2(A).$$

The statement holds by noticing that $\sum_{j=1}^{n} \|A_j\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} a_{ij}^2$. □

2 Singular Value Decomposition

We can think of $A \in \mathbb{R}^{n\times d}$ as a linear transformation taking a vector v_1 in its row space to a vector $u_1 = Av_1$ in its column space. Many applications require to find an orthogonal basis for the row space and transform it into an orthogonal basis for the column space: $Av_i = \sigma_i u_i$. The heart of the problem is to find $v_1, \ldots, v_r$ for the row space of A for which

$$
\begin{aligned}
A[v_1, v_2, \ldots, v_r] &= [\sigma_1 u_1, \sigma_2 u_2, \ldots, \sigma_r u_r] \\
&= [u_1, u_2, \ldots, u_r] \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix}.
\end{aligned} \tag{1}
$$

Then, it is easy to see that the left and right-singular vectors $u_i = \frac{1}{\sigma_i} Av_i$, v_i, and their associated singular values σ_i satisfy (1). With these vectors u_is, v_is, and the singular values σ_is, we can write A in matrix notation as

$$A = UDV^{\intercal},$$

where u_i is the i-th column of U, $v_i^{\intercal}$ is the i-th row of $V^{\intercal}$, and D is the diagonal matrix with σ_i as the i-th entry on its diagonal. This factorisation of A in the form of $UDV^{\intercal}$ is called *Singular value decomposition.* It is easy to check that

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^{\intercal}.$$