Singular Value Decomposition and Applications

Uri Shaham

March 4, 2024

1 Introduction

Definition 1.1 (SVD)**.** *Let $A \in \mathbb{R}^{n\times m}$ be a real-valued matrix. The singular value decomposition (SVD) of A is a matrix factorization*

$$A = U\Sigma V^T,$$

where U is $n \times n$ orthogonal matrix (i.e., $U^TU = I_{n\times n}$), V is $m \times m$ orthogonal matrix and Σ is $n \times m$ "diagonal" matrix (i.e., $\Sigma_{ij} = 0$ for $i \neq j$, with nonnegative entries.

Observation 1.2. *$U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ where $r = \min\{n, m\}$, where $\sigma_i := \Sigma_{ii}$.*

Remark 1.3. *Since only terms corresponding to nonzero singular values matter in the SVD of a $n \times m$ matrix A, it is often convenient to include only the corresponding terms in the SVD, i.e., viewing the matrix U as $n \times r$, Σ as $r \times r$ and V as $m \times r$. This is called the "compact" or "reduced" representation of the SVD.*

Remark 1.4. *Without loss of generality, it is convenient to assume that the singular values are decreasingly ordered, i.e., $\sigma_i \geq \sigma_j$ for $i < j$.*

Observation 1.5. *When $n = m$, A can be viewed as an operator from $\mathbb{R}^n$ to $\mathbb{R}^n$, acting on any vector x by rotation (possibly with reflection), axis rescaling and another rotation.*

1.1 Existence and uniqueness of SVD

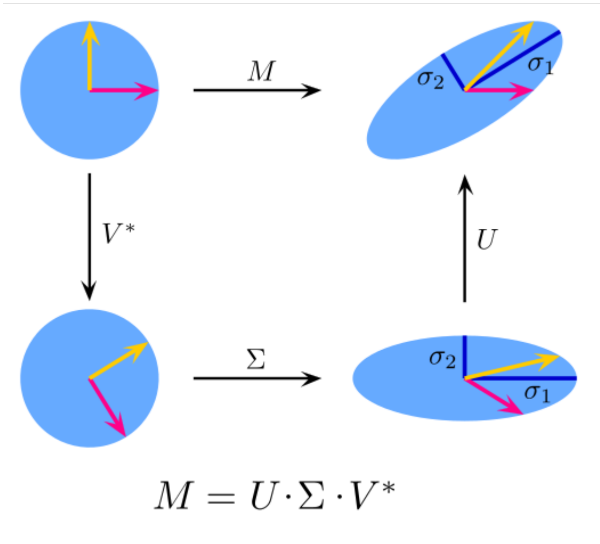
Theorem 1.6 (Existence of SVD)**.** *Any matrix $A \in \mathbb{R}^{n\times m}$ has a SVD.*

Proof. The matrix A^TA is symmetric (clearly) and positive semi-definite (to see this, assume that $\lambda < 0$ is an eigenvalue and let x be the corresponding eigenvector. Then

$$\sum_i (Ax)_i^2 = (Ax)^T(Ax) = x^TA^TAx < 0,$$

which is a contradiction.). Then A^TA has an eigendecomposition $A^TA = V\Lambda V^T$ with real (and orthogonal) eigenvectors and non-negative eigenvalues. Let $r = \text{rank}(A^TA)$. Wlog, assume that $\lambda_1 \geq \lambda_2, \ldots \geq \lambda_r > 0$ and $\lambda_{r+1} = \ldots = \lambda_m = 0$. Set $\sigma_i = \sqrt{\lambda_i}$, for $i = 1, \ldots, r$. Define $u_i = \frac{Av_i}{\sigma_i}$ for $i = 1, \ldots, r$. Then $u_1, \ldots, u_r$ are orthonormal:

$$u_i^T u_j = \frac{(v_i^TA^T)Av_j}{\sigma_i\sigma_j} = \frac{v_i^T(A^TAv_j)}{\sigma_i\sigma_j} = \frac{v_i^T(\lambda_j v_j)}{\sigma_i\sigma_j} = \delta_{ij}.$$



Then $U = AV\Sigma^{-1}$, so

$$U\Sigma V^T = AV\Sigma^{-1}\Sigma V^T = A.$$

□

Theorem 1.7. *Let $A = U\Sigma V^T$ Then Σ is uniquely determined.*

Proof. This follows from the fact that the singular values of A are the square roots of the eigenvalues of $A^T A$, which are uniquely determined, up to order (being the roots of the characteristic polynomial of $A^T A$).

□

1.2 Power iteration

Observation 1.8. *Let $A = U\Sigma V^T$ and let $B = A^T A$. Then $B = V\Sigma^2 V^T$ and more generally, $B^k = V\Sigma^{2k}V^T$. Note that Σ^{2k} is diagonal with entries σ_i^{2k}.*

Assume that $\sigma_1 > \sigma_2$. Then for k large enough $\sigma_1^{2k} \gg \sigma_2^{2k}$, hence

$$B^k = \sum_i \sigma_i^{2k} v_i v_i^T \approx \sigma_1^{2k} v_1 v_1^T.$$

Let x be an arbitrary vector with nonzero component in the direction of v_1, i.e., $x = \sum_{i=1}^m \alpha_i v_i$, and $\alpha_1 \neq 0$. Then for sufficiently large k, $B^k x \approx \sigma_1^{2k}\alpha_1 v_1$, i.e., $B^k x$ is approximately in the direction of v_1, so v_1, $\frac{B^k x}{\|B^k x\|} \to v_1$ as $k \to \infty$. This gives an approach to find v_1:

- Starting from any vector x_0 not orthogonal to v_1 (a random vector would typically work):
- Repeat until $\|x_t - x_{t-1}\|_2 \leq \epsilon$:
    1. $x_t \leftarrow Bx_{t-1}$

2. Normalize x to be of unit length, i.e., $x_t \leftarrow \frac{x_t}{\|x_t\|}$.

In the homework you will extend this method to subsequent singular vectors.

2 Applications

2.1 Low rank approximation

Definition 2.1 (spectral norm)**.** *Let $A = U\Sigma V^T = \sum_{i=1}^r u_i v_i^T$ be $n \times m$ matrix. The spectral norm (also known as the operator norm) of A is defined as its largest singular value, i.e., $\|A\|_2 := \sigma_1$.*

Theorem 2.2 (spetral norm is matrix 2-norm.)**.** $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$

This will be proven next week,.

Theorem 2.3 (Eckart-Young, 1936)**.** *The best rank k approximation of A in spectral norm is $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$.*

Proof. First, note that $\|A - A_k\|_2 = \|\sum_{i=k+1}^r \sigma_i u_i v_i^T\|_2 = \sigma_{k+1}$. Let B_k be any $n \times m$ rank-k matrix, i.e., $B_k = XY^T$, where X and Y have k columns each. Since Y has k columns, there is a vector $w \in \text{span}\{v_1, \ldots, v_{k+1}\}$ which is orthogonal to any column in Y, i.e., $w := \sum_{j=1}^{k+1} \gamma_j v_j$ gives $Y^T w = 0$. Then $B_k w = 0$. Wlog $\|w\| = 1$, i.e., $\sum_{i=1}^{k+1} \gamma^2 = 1$ (by Pythagoras). Hence we have

$$
\begin{aligned}
\|A - B_k\|_2^2 &\geq \|(A - B_k)w\|_2^2 \text{ (due to theorem 2.2)} \\
&= \|Aw\|_2^2 \\
&= \left\| \sum_i \sigma_i u_i v_i^T \left(\sum_{j=1}^{k+1} \gamma_j v_j \right) \right\|_2^2 \\
&= \left\| \sum_{i=1}^{k+1} \sigma_i \gamma_i u_i \right\|_2^2 \\
&= \sum_{i=1}^{k+1} \sigma_i^2 \gamma_i^2 \text{ (as the above is a norm of a vector, expanded in the basis } \{u_1, \ldots, u_m\}) \\
&\geq \sigma_{k+1} \sum_{i=1}^{k+1} \gamma_i^2 \\
&= \sigma_{k+1} \\
&= \|A - A_k\|.
\end{aligned}
$$

□

2.2 Pseudo inverse

Definition 2.4. *The pseudo inverse of a full rank $n \times d$ matrix (with $n \geq d$) $X = U\Sigma V^T$ is*

$$X^\dagger := (X^T X)^{-1} X^T = V\Sigma^{-2} V^T V \Sigma U^T = V \Sigma^{-1} U^T.$$

Remark 2.5. *Note that if X is not full rank, or if $n < d$, $X^T X$ is not invertible.*

From linear regression, we know that pseudo inverse can be used to solve least squares problems as follows. Let $(X \in \mathbb{R}^{n\times d}, y \in \mathbb{R}^n)$ *be a training set of regression data, and let* $\beta = X^\dagger y$*. Then* β *minimizes the least squares prediction error, i.e.,*

$$\beta = \arg\min_{b\in\mathbb{R}^d} \|Xb - y\|^2.$$

In the more general case, we write $X^\dagger = V\Sigma^\dagger U^T$*, where* $\Sigma^\dagger$ *is obtained from* Σ *by replacing all nonzero singular values by their reciprocals.*

2.3 Matrix square root

Let A be a symmetric $n \times n$ PSD matrix with SVD $A = V\Sigma V^T$ (in the homework you will be asked to prove $U = V$ whenever A is symmetric and PSD). Let $\Sigma^{\frac{1}{2}}$ be $\mathrm{diag}(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_n})$. Then for $B = V\Sigma^{\frac{1}{2}}V^T$ we have $BB = A$.

2.4 Sampling from multivariate normal distribution

Let K be a $d\times d$ covariance matrix (note that in particular, it is symmetric and PSD). To sample $y \in \mathbb{R}^d$ from a $\mathcal{N}(\mu, K)$ normal distribution:

1. For $i = 1, \dots d$ sample $x_i \sim \mathcal{N}(0, 1)$.
2. Set $y = \mu + K^{\frac{1}{2}}x$.

Alternatively, let $K = V\Sigma V^T$ be the SVD of K.

1. For $i = 1, \dots d$ sample $x_i \sim \mathcal{N}((V^T\mu)_i, \sigma_i)$ - easy (why?).
2. Set $y = Vx$.

2.5 PCA

Let X be a $n\times d$ matrix with mean-centered columns, representing n data points in d dimensions. Then the sample covariance matrix is X^TX (up to constant multiplication). In PCA, the principal directions are the eigenvectors V of the covariance matrix. If $X = U\Sigma V^T$, the covariance is $V\Sigma^2V^T$, therefore the PCA embedding is $XV = U\Sigma$. The reconstruction is given by multiplying the embedding from the right by V^T, i.e., $U\Sigma V^T = X$. Reconstruction from fewer terms therefore amounts to low-rank approximation of X.