THE SINGULAR VALUE DECOMPOSITION

SYNOPSIS. The linear algebra framework in which matrix-vector multiplication represents a linear combination and also corresponds to evaluation of a linear mapping has been shown to be complete through the FTLA. In order to effectively solve the main problems of linear algebra within this framework, bases must be constructed for the fundamental subspaces of a matrix. In order for the bases to be computationally efficient they should be orthonormal. The existence of such basis sets is guaranteed by the singular value decomposition theorem.

1. Orthogonal matrices

The fundamental theorem of linear algebra partitions the domain and codomain of a linear mapping $\boldsymbol{f}: U \longrightarrow V$. For real vectors spaces $U = \mathbb{R}^n$, $V = \mathbb{R}^m$ the partition properties are stated in terms of spaces of the associated matrix $\boldsymbol{A}$ as

$$C(\boldsymbol{A}) \oplus N(\boldsymbol{A}^T) = \mathbb{R}^m \quad C(\boldsymbol{A}) \perp N(\boldsymbol{A}^T) \quad C(\boldsymbol{A}^T) \oplus N(\boldsymbol{A}) = \mathbb{R}^n \quad C(\boldsymbol{A}^T) \perp N(\boldsymbol{A}) \; .$$

The dimension of the column and row spaces $r = \dim C(\boldsymbol{A}) = \dim C(\boldsymbol{A}^T)$ is the rank of the matrix, $n - r$ is the nullity of $\boldsymbol{A}$, and $m - r$ is the nullity of A^T. A infinite number of bases could be defined for the domain and codomain. It is of great theoretical and practical interest to define bases with properties that facilitate insight or computation.

The above partitions of the domain and codomain are orthogonal, and suggest searching for orthogonal bases within these subspaces. Introduce a matrix representation for the bases

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \end{bmatrix} \in \mathbb{R}^{m \times m}, \boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_n \end{bmatrix} \in \mathbb{R}^{n \times n},$$

with $C(\boldsymbol{U}) = \mathbb{R}^m$ and $C(\boldsymbol{V}) = \mathbb{R}^n$. Orthogonality between columns $\boldsymbol{u}_i$, $\boldsymbol{u}_j$ for $i \neq j$ is expressed as $\boldsymbol{u}_i^T \boldsymbol{u}_j = 0$. For $i = j$, the inner product is positive $\boldsymbol{u}_i^T \boldsymbol{u}_i > 0$, and since scaling of the columns of $\boldsymbol{U}$ preserves the spanning property $C(\boldsymbol{U}) = \mathbb{R}^m$, it is convenient to impose $\boldsymbol{u}_i^T \boldsymbol{u}_i = 1$. Such behavior is concisely expressed as a matrix product

$$\boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}_m,$$

with $\boldsymbol{I}_m$ the identity matrix in $\mathbb{R}^m$. Expanded in terms of the column vectors of $\boldsymbol{U}$ the first equality is

$$\begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \end{bmatrix}^T \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \\ \vdots \\ \boldsymbol{u}_m^T \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^T \boldsymbol{u}_1 & \boldsymbol{u}_1^T \boldsymbol{u}_2 & \dots & \boldsymbol{u}_1^T \boldsymbol{u}_m \\ \boldsymbol{u}_2^T \boldsymbol{u}_1 & \boldsymbol{u}_2^T \boldsymbol{u}_2 & \dots & \boldsymbol{u}_2^T \boldsymbol{u}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{u}_m^T \boldsymbol{u}_1 & \boldsymbol{u}_m^T \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m^T \boldsymbol{u}_m \end{bmatrix} = \boldsymbol{I}_m.$$

It is useful to determine if a matrix $\boldsymbol{X}$ exists such that $\boldsymbol{U}\boldsymbol{X} = \boldsymbol{I}_m$, or

$$\boldsymbol{U}\boldsymbol{X} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix}.$$

The columns of $\boldsymbol{X}$ are the coordinates of the column vectors of $\boldsymbol{I}_m$ in the basis $\boldsymbol{U}$, and can readily be determined

$$\boldsymbol{U}\boldsymbol{x}_j = \boldsymbol{e}_j \Longrightarrow \boldsymbol{U}^T \boldsymbol{U}\boldsymbol{x}_j = \boldsymbol{U}^T \boldsymbol{e}_j \Longrightarrow \boldsymbol{I}_m \boldsymbol{x}_j = \begin{bmatrix} \boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \\ \vdots \\ \boldsymbol{u}_m^T \end{bmatrix} \boldsymbol{e}_j \Longrightarrow \boldsymbol{x}_j = (\boldsymbol{U}^T)_j,$$

where $(\boldsymbol{U}^T)_j$ is the j^{th} column of $\boldsymbol{U}^T$, hence $\boldsymbol{X} = \boldsymbol{U}^T$, leading to

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} = \boldsymbol{U}\boldsymbol{U}^T.$$

Note that the second equality

$$\begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \end{bmatrix}\begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \end{bmatrix}^T = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_m \end{bmatrix}\begin{bmatrix} \boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \\ \vdots \\ \boldsymbol{u}_m^T \end{bmatrix} = \boldsymbol{u}_1\boldsymbol{u}_1^T + \boldsymbol{u}_2\boldsymbol{u}_2^T + \cdots + \boldsymbol{u}_m\boldsymbol{u}_m^T = \boldsymbol{I}$$

acts as normalization condition on the matrices $\boldsymbol{U}_j = \boldsymbol{u}_j\boldsymbol{u}_j^T$.

DEFINITION. *A square matrix $\boldsymbol{U}$ is said to be orthogonal if $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$.*

2. Intrinsic basis of a linear mapping

Given a linear mapping $\boldsymbol{f}: U \longrightarrow V$, expressed as $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, the simplest description of the action of $\boldsymbol{A}$ would be a simple scaling, as exemplified by $\boldsymbol{g}(\boldsymbol{x}) = a\boldsymbol{x}$ that has as its associated matrix $a\boldsymbol{I}$. Recall that specification of a vector is typically done in terms of the identity matrix $\boldsymbol{b} = \boldsymbol{I}\boldsymbol{b}$, but may be more insightfully given in some other basis $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{I}\boldsymbol{b}$. This suggests that especially useful bases for the domain and codomain would reduce the action of a linear mapping to scaling along orthogonal directions, and evaluate $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ by first re-expressing $\boldsymbol{y}$ in another basis $\boldsymbol{U}$, $\boldsymbol{U}\boldsymbol{s} = \boldsymbol{I}\boldsymbol{y}$ and re-expressing $\boldsymbol{x}$ in another basis $\boldsymbol{V}$, $\boldsymbol{V}\boldsymbol{r} = \boldsymbol{I}\boldsymbol{x}$. The condition that the linear operator reduces to simple scaling in these new bases is expressed as $s_i = \sigma_i r_i$ for $i = 1,\dots,\min(m,n)$, with σ_i the scaling coefficients along each direction which can be expressed as a matrix vector product $\boldsymbol{s} = \boldsymbol{\Sigma}\boldsymbol{r}$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ is of the same dimensions as $\boldsymbol{A}$ and given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_r & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix}.$$

Imposing the condition that $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal leads to

$$\boldsymbol{U}\boldsymbol{s} = \boldsymbol{y} \Longrightarrow \boldsymbol{s} = \boldsymbol{U}^T\boldsymbol{y}, \boldsymbol{V}\boldsymbol{r} = \boldsymbol{x} \Longrightarrow \boldsymbol{r} = \boldsymbol{V}^T\boldsymbol{x},$$

which can be replaced into $\boldsymbol{s} = \boldsymbol{\Sigma}\boldsymbol{r}$ to obtain

$$\boldsymbol{U}^T\boldsymbol{y} = \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x} \Longrightarrow \boldsymbol{y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}.$$

From the above the orthogonal bases $\boldsymbol{U}, \boldsymbol{V}$ and scaling coefficients $\boldsymbol{\Sigma}$ that are sought must satisfy $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. The SVD theorem states that the matrix factors $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ do indeed exist.

THEOREM. *Every matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has a singular value decomposition (SVD)*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

with properties:

1. $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ *is an orthogonal matrix,* $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_m$;
2. $\boldsymbol{V} \in \mathbb{R}^{m\times m}$ *is an orthogonal matrix,* $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_n$;
3. $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ *is diagonal,* $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_p)$, $p = \min(m, n)$, *and* $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_p \geqslant 0$.

The scaling coefficients σ_j are called the *singular values* of $\boldsymbol{A}$. The columns of $\boldsymbol{U}$ are called the *left singular vectors*, and those of $\boldsymbol{V}$ are called the *right singular vectors*. Carrying out computation of the matrix products

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \ldots & \boldsymbol{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_r^T \\ \boldsymbol{v}_{r+1}^T \\ \vdots \\ \boldsymbol{v}_n^T \end{bmatrix} \Longrightarrow$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \ldots & \boldsymbol{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1\boldsymbol{v}_1^T \\ \sigma_2\boldsymbol{v}_2^T \\ \vdots \\ \sigma_r\boldsymbol{v}_r^T \\ \vdots \\ 0 \end{bmatrix}$$

leads to a representation of $\boldsymbol{A}$ as a sum

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T,$$

with $r \leqslant \min(m, n)$. Written out in full, the above sum is

$$\boldsymbol{A} = \sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^T + \sigma_2\boldsymbol{u}_2\boldsymbol{v}_2^T + \cdots + \sigma_r\boldsymbol{u}_r\boldsymbol{v}_r^T.$$

Each product $\boldsymbol{u}_i\boldsymbol{v}_i^T$ is a matrix of rank one, and is called a rank-one update. Truncation of the above sum to p terms leads to an approximation of $\boldsymbol{A}$

$$\boldsymbol{A} \cong \boldsymbol{A}_p = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T.$$

In very many cases the singular values exhibit rapid decay, $\sigma_1 \gg \sigma_2 \gg \cdots$, such that the approximation above is an accurate representation of the matrix $\boldsymbol{A}$ for $p \ll r$.

The singular vector matrices $\boldsymbol{U}, \boldsymbol{V}$ specify the intrinsic directions within $\mathbb{R}^m, \mathbb{R}^n$ along which the matrix $\boldsymbol{A}$ acts as a simple scaling transformation. For example, applying the linear mapping to the $\boldsymbol{v}_1$ vector, $\boldsymbol{f}(\boldsymbol{v}_1) = \boldsymbol{A}\boldsymbol{v}_1$, leads to

$$\boldsymbol{A}\boldsymbol{v}_1 = \left(\sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \right) \boldsymbol{v}_1 = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i (\boldsymbol{v}_i^T \boldsymbol{v}_1) = \sigma_1 \boldsymbol{u}_1.$$

Since $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$, the above states that the input direction most amplified by the $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ mapping is $\boldsymbol{v}_1$ and the result is the vector $\sigma_1 \boldsymbol{u}_1$. The two-norm of $\boldsymbol{v}_1$ is equal to one and that of $\sigma_1 \boldsymbol{u}_1$ is σ_1. The conclusion is that σ_1 is the maximal amplification factor in the two-norm

$$\sigma_1 = \max_{\|\boldsymbol{x}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{x}\|_2,$$

and the above satisfies the properties of a norm over matrices leading to the definition

$$\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{x}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{x}\|_2.$$

The largest singular value is thus the two-norm of a matrix.

3. SVD solution of linear algebra problems

The SVD can be used to solve common problems within linear algebra.

Linear systems. To change from vector coordinates $\boldsymbol{b}$ in the canonical basis $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ to coordinates $\boldsymbol{x}$ in some other basis $\boldsymbol{A} \in \mathbb{R}^{m \times m}$, a solution to the equation $\boldsymbol{I}\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ can be found by the following steps.

1. Compute the SVD, $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{A}$;
2. Find the coordinates of $\boldsymbol{b}$ in the orthogonal basis $\boldsymbol{U}$, $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{b}$;
3. Scale the coordinates of $\boldsymbol{c}$ by the inverse of the singular values $y_i = c_i/\sigma_i$, $i = 1, \ldots, m$, such that $\Sigma\, \boldsymbol{y} = \boldsymbol{c}$ is satisfied;
4. Find the coordinates of $\boldsymbol{y}$ in basis $\boldsymbol{V}^T$, $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{y}$.

Least squares. In the above $\boldsymbol{A}$ was assumed to be a basis, hence $r = \text{rank}(\boldsymbol{A}) = m$. If columns of $\boldsymbol{A}$ do not form a basis, $r < m$, then $\boldsymbol{b} \in \mathbb{R}^m$ might not be reachable by linear combinations within $C(\boldsymbol{A})$. The closest vector to $\boldsymbol{b}$ in the 2-norm is however found by the same steps, with the simple modification that in Step 3, the scaling is carried out only for non-zero singular values, $y_i = c_i/\sigma_i$, $i = 1, \ldots, r$.

The pseudo-inverse. From the above, finding either the solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{I}\boldsymbol{b}$ or the best approximation possible if $\boldsymbol{A}$ is not of full rank, can be written as a sequence of matrix multiplications using the SVD

$$(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)\,\boldsymbol{x} = \boldsymbol{b} \Longrightarrow \boldsymbol{U}\,(\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}) = \boldsymbol{b} \Longrightarrow (\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}) = \boldsymbol{U}^T\boldsymbol{b} \Longrightarrow \boldsymbol{V}^T\boldsymbol{x} = \boldsymbol{\Sigma}^+\boldsymbol{U}^T\boldsymbol{b} \Longrightarrow \boldsymbol{x} = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^T\boldsymbol{b},$$

where the matrix $\boldsymbol{\Sigma}^+ \in \mathbb{R}^{n \times m}$ (notice the inversion of dimensions) is defined as a matrix with elements σ_i^{-1} on the diagonal, and is called the pseudo-inverse of $\boldsymbol{\Sigma}$. Similarly the matrix

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^T$$

that allows stating the solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ simply as $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b}$ is called the *pseudo-inverse* of $\boldsymbol{A}$. Note that in practice $\boldsymbol{A}^+$ is not explicitly formed. Rather the notation $\boldsymbol{A}^+$ is simply a concise reference to carrying out steps 1-4 above.