The Singular Value Decomposition of a (Real) Matrix

We have discussed several ways to factorize matrices, including

- **LU factorization**: Any m-by-n matrix $\mathbf{A}$ is eligible. If $\mathbf{A}$ can be put into RREF without row swaps, then $\mathbf{A} = \mathbf{LU}$. Otherwise, there is a permutation matrix $\mathbf{P}$ for which $\mathbf{PA} = \mathbf{LU}$. This sort of decomposition for $\mathbf{A}$ is ideally suited to Gaussian elimination—i.e., to solve $\mathbf{Ax} = \mathbf{b}$.
- **QR factorization**: The Gram-Schmidt process can be used with any m-by-n matrix $\mathbf{A}$ to produce an m-by-n matrix $\mathbf{Q}$ with orthonormal columns and an upper triangular n-by-n matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$. This sort of decomposition for $\mathbf{A}$ is ideally suited to finding the projections in subspaces, and to finding a least-squares solution when $\mathbf{Ax} = \mathbf{b}$ is inconsistent.
- **Diagonalization**: If the square (n-by-n) matrix $\mathbf{A}$ is diagonalizable, then there exists a matrix $\mathbf{X}$ of eigenvectors such that $\mathbf{A} = \mathbf{X\Lambda X}^{-1}$. When $\mathbf{A}$ is symmetric, we know can be chosen to form an orthonormal basis of $\mathbf{R}^n$. In that case, $\mathbf{X}$ is an orthogonal matrix, $\mathbf{X}^{-1} = \mathbf{X}^{\mathrm{T}}$, and $\mathbf{A} = \mathbf{X\Lambda X}^{\mathrm{T}}$. This sort of factorization is nice for finding powers of $\mathbf{A}$, and for solving systems of 1st-order linear differential equations.

Every decomposition has settings in which it is convenient to use. The singular value decomposition is no exception, as examples involving image compression will demonstrate.

The SVD is available for any real m-by-n matrix $\mathbf{A}$. It says

$$\mathbf{A} \;=\; \mathbf{U\Sigma V}^{\mathrm{T}},$$

where $\mathbf{\Sigma}$ is a “diagonal” m-by-n matrix, $\mathbf{U}$ is an orthogonal matrix whose columns form an orthonormal basis for $\mathbb{R}^m$, and $\mathbf{V}$ is an orthogonal matrix whose columns form an orthonormal basis for $\mathbb{R}^n$, Here is how it arises.

1. We start with an m-by-n matrix $\mathbf{A}$ whose rank is r. We can show (and have done so, I think) that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{AA}^{\mathrm{T}}$ are symmetric and positive semidefinite. In particular, all eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are real and $\geqslant 0$. We can index these eigenvalues σ_i^2 in descending order, so that

$$\sigma_1^2 \geqslant \sigma_2^2 \geqslant \ldots \geqslant \sigma_r^2 > \sigma_{r+1}^2 = \cdots = \sigma_n^2 = 0. \tag{1}$$

   The Spectral Theorem says we can choose corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ so that they form an orthonormal basis of $\mathbb{R}^n$, and by constructing $\mathbf{V}$ so that its j^{th} column is $\mathbf{v}_j$, we diagonalize $\mathbf{A}^{\mathrm{T}}\mathbf{A}$:

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} \;=\; \mathbf{V}\begin{bmatrix} \begin{matrix}\sigma_1^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_r^2\end{matrix} & \mathbf{0} \\ \mathbf{0} & \begin{matrix}0 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 0\end{matrix}\end{bmatrix}\mathbf{V}^{\mathrm{T}}.$$

In fact, the $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ all lie in the row space of $\mathbf{A}$ (Showing this is a good exercise!) and, being independent and plentiful in the correct number, form a basis for $\mathrm{Col}(\mathbf{A}^{\mathrm{T}})$. The rest of the columns, $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ go with the eigenvalue 0, making them a basis for $\mathrm{Null}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \mathrm{Null}(\mathbf{A})$.

Note: We wrote the eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ as squares, which we could do knowing they were real and nonnegative. In what follows, we will refer to the unsquared values σ_j, by which we mean the nonnegative square root of σ_j^2.

2. Next, for $j = 1, \ldots, r$, we can define

$$\mathbf{u}_j \;:=\; \frac{1}{\sigma_j}\mathbf{A}\mathbf{v}_j.$$

Then for $1 \leqslant j, k \leqslant r$,

$$\begin{aligned}\langle \mathbf{u}_j, \mathbf{u}_k\rangle \;&=\; \left\langle \frac{1}{\sigma_j}A\mathbf{v}_j, \frac{1}{\sigma_k}A\mathbf{v}_k\right\rangle \;=\; \frac{1}{\sigma_j\sigma_k}\langle \mathbf{A}\mathbf{v}_j, \mathbf{A}\mathbf{v}_k\rangle \;=\; \frac{1}{\sigma_j\sigma_k}\langle \mathbf{v}_j, \mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{v}_k\rangle \\ &=\; \frac{1}{\sigma_j\sigma_k}\langle \mathbf{v}_j, \sigma_k^2\mathbf{v}_k\rangle \;=\; \frac{\sigma_k}{\sigma_j}\langle \mathbf{v}_j, \mathbf{v}_k\rangle \;=\; \begin{cases} 0, & \text{if } j \neq k, \\ 1, & \text{if } j = k,\end{cases}\end{aligned}$$

since the $\mathbf{v}_k$ are orthonormal. Thus, $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal set.

3. By construction, the $\mathbf{u}_j$ satisfy

$$\mathbf{A}\mathbf{v}_j \;=\; \sigma_j\mathbf{u}_j, \qquad \text{for} \qquad j = 1, \ldots, r.$$

Moreover,

$$\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{u}_j \;=\; \mathbf{A}\mathbf{A}^{\mathrm{T}}\left(\frac{1}{\sigma_j}\mathbf{A}\mathbf{v}_j\right) \;=\; \frac{1}{\sigma_j}\mathbf{A}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{v}_j\right) \;=\; \frac{1}{\sigma_j}\mathbf{A}\left(\sigma_j^2\mathbf{v}_j\right) \;=\; \sigma_j^2 \cdot \frac{1}{\sigma_j}\mathbf{A}\mathbf{v}_j \;=\; \sigma_j^2\mathbf{u}_j,$$

which shows these $\mathbf{u}_j$, $j = 1, \ldots, r$, are eigenvectors of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, corresponding to eigenvalues $\sigma_1^2, \ldots, \sigma_r^2$. The $\mathbf{u}_j$ reside in $\mathrm{Col}(\mathbf{A})$, and form a basis for it, since $\mathrm{rank}(\mathbf{A}\mathbf{A}^{\mathrm{T}}) = \mathrm{rank}(\mathbf{A}^{\mathrm{T}}) = \mathrm{rank}(\mathbf{A})$. All the remaining eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ are zero. Thus, in the diagonalization of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$,

$$\mathbf{A}\mathbf{A}^{\mathrm{T}} \;=\; \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}},$$

the eigenvalues are the same as for $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, and can be arranged largest to smallest as in Expression (1). Moreover, the first r columns of $\mathbf{U}$ can be taken, in order, as $\mathbf{u}_1, \ldots, \mathbf{u}_r$.

The rest of the columns come from the eigenspace of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ corresponding to eigenvalue 0—that is, from $\mathrm{Null}(\mathbf{A}\mathbf{A}^{\mathrm{T}}) = \mathrm{Null}(\mathbf{A}^{\mathrm{T}})$, and can be made to complete the orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ of $\mathbb{R}^m$.