# SINGULAR VALUE DECOMPOSITION

Vega Federico Gaspar

## 1 INTRODUCTION

The singular value decomposition (SVD) is among the most important matrix factorizations of the computational era. The SVD provides a numerically stable matrix decomposition that can be used for a variety of purposes. For example, to obtain a low-rank approximations to matrices and to perform pseudo-inverses of non-square matrices to find the solution of a linear system of equations. Another important use of the SVD is as the underlying algorithm of principal component analysis (PCA), where high-dimensional data is decomposed into its most statistically descriptive factors. SVD/PCA has been applied to a wide variety of problems in science and engineering. The SVD provides a systematic way to determine a low-dimensional approximation to high-dimensional data in terms of dominant patterns. The SVD is numerically stable and provides a hierarchical representation of the data in terms of a new coordinate system defined by dominant correlations within the data. Moreover, the SVD is guaranteed to exist for any matrix, unlike the eigendecomposition.

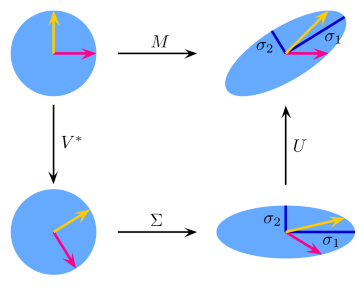
## 2 INTUITIVE INTERPRETATIONS

### 2.1 ROTATION, COORDINATE SCALING, AND REFLECTION

Any invertible linear transformation $x \to Ax$ can be expressed as a composition of three geometrical transformations: a rotation or reflection ($\mathbf{V}^\dagger$), followed by a coordinate-by-coordinate scaling ($\mathbf{\Sigma}$), followed by another rotation or reflection ($\mathbf{U}$). In particular, if $\mathbf{X}$ (the matrix representation of $A$) has a positive determinant, then $\mathbf{U}$ and $\mathbf{V}^*$ can be chosen to be both reflections, or both rotations. If the determinant is negative, exactly one of them will have to be a reflection. If the determinant is zero, each can be independently chosen to be of either type.

If the matrix $\mathbf{X}$ is real but not square, namely $m \times n$ with $m \neq n$, it can be interpreted as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $\mathbf{U}$ and $\mathbf{V}^\dagger$ can be chosen to be rotations of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively; and $\mathbf{\Sigma}$, besides scaling the first $\min\{m, n\}$ coordinates, also extends the vector with zeros, i.e. removes trailing coordinates, so as to turn $\mathbb{R}^n$ into $\mathbb{R}^m$.

### 2.2 SINGULAR VALUES AS SEMI-AXES OF AN ELLIPSE OR ELLIPSOID




As shown in the figure, the singular values can be interpreted as the magnitude of the semiaxes of an ellipse in 2D. This concept can be generalized to n-dimensional Euclidean space, with the singular values of any $n \times n$ square matrix being viewed as the magnitude of the semiaxis of an n-dimensional ellipsoid. Similarly, the singular values of any $m \times n$ matrix can be viewed as the magnitude of the semiaxis of an n-dimensional ellipsoid in m-dimensional space, for example as an ellipse in a (tilted) 2D plane in a 3D space. Singular values encode magnitude of the semiaxis, while singular vectors encode direction.

### 2.3 THE COLUMNS OF U AND V FORMS AN ORTHONORMAL BASES

Since $\mathbf{U}$ and $\mathbf{V}^\dagger$ are unitary, the columns of each of them form a set of orthonormal vectors, which can be regarded as basis vectors. In other words, the columns $\mathbf{U}_1, ..., \mathbf{U}_m$ of $\mathbf{U}$ yield an orthonormal basis of $\mathbb{R}^m$ and the columns $\mathbf{V}_1, ..., \mathbf{V}_n$ of $\mathbf{U}$ yield an orthonormal basis of $\mathbb{R}^n$. If the columns of $\mathbf{X}$ are spatial measurements in time, then $\mathbf{U}$ encode spatial patterns, and $\mathbf{V}$ encode temporal patterns. Also, as a consequence of unitarity, solving a system of equations involving $\mathbf{U}$ or $\mathbf{V}$ is as simple as multiplication by the

transpose, which scales as $O\left(n^2\right)$, as opposed to traditional methods for the generic inverse, which scale as $O\left(n^3\right)$.

The linear transformation $\mathbf{X}$ has a particularly simple description with respect to these orthonormal bases: we have

$$\mathbf{X}(\mathbf{V}_i) = \sigma_i \mathbf{U}_i, \qquad i = 1, \ldots, \min(m, n),$$

where $\sigma_i$ is the i-th diagonal entry of $\mathbf{\Sigma}$, and $\mathbf{X}(\mathbf{V}_i) = 0$ for i > min(m,n).

The geometric content of the SVD theorem can thus be summarized as follows: for every linear map $X : K^n \to K^m$ one can find orthonormal bases of $K^n$ and $K^m$ such that $X$ maps the i-th basis vector of $K^n$ to a non-negative multiple of the i-th basis vector of $K^m$, and sends the left-over basis vectors to zero. With respect to these bases, the map $X$ is therefore represented by a diagonal matrix with non-negative real diagonal entries.

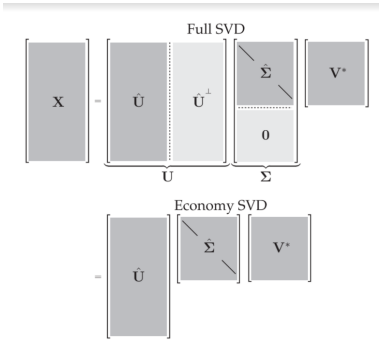
## 3 Definition of SVD

Generally, we are interested in analyzing a large data set $\mathbf{X} \in \mathbb{C}^{n \times m}$, where the columns $x_k \in \mathbb{C}^n$ may be measurements from simulations or experiments. For example, columns may represent images that have been reshaped into column vectors with as many elements as pixels in the image. The column vectors may also represent the state of a physical system that is evolving in time, such as the fluid velocity at a set of discrete points. The index $k$ is a label indicating the k-th distinct set of measurements. For many of the examples, $\mathbf{X}$ will consist of a time-series of data, and $x_k = x(k\Delta t)$. Often the state-dimension n is very large, on the order of millions or billions of degrees of freedom. The columns are often called snapshots, and the numbers of columns $m$ is the number of snapshots in $X$. For many systems $n >> m$, resulting in a tall-skinny matrix, as opposed to a short-fat matrix when $n << m$.

The SVD is a unique matrix decomposition that exists for every complex-valued matrix $\mathbf{X} \in \mathbb{C}^{n \times m}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger \tag{1}$$

where $\mathbf{U} \in \mathbb{C}^{n \times m}$ and $\mathbf{V} \in \mathbb{C}^{n \times m}$ are unitary matrices with orthonormal columns, and $\Sigma \in \mathbb{R}^{n \times m}$ is a matrix with real, nonnegative entries on the diagonal and zeros off the diagonal. When $n \geq m$, the matrix has at most $m$ nonzero elements on the diagonal. Therefore, it is possible to **exactly** represent $\mathbf{X}$ using the economy SVD:




The columns of $\hat{\mathbf{U}}^\perp$ span a vector space that is complementary and orthogonal to that spanned by $\hat{\mathbf{U}}$ . The columns of $\mathbf{U}$ are called left singular vectors of $\mathbf{X}$ and the columns of $\mathbf{V}$ are right singular vectors. The diagonal elements of $\hat{\mathbf{\Sigma}} \in \mathbb{C}^{m \times m}$ are called singular values and they are ordered from largest to smallest. The rank of $\mathbf{X}$ is equal to the number of nonzero singular values.

### 3.1 DOMINANT CORRELATIONS

Using that $\mathbf{U}$ and $\mathbf{V}$ are unitary it is easy to prove that:

$$\begin{cases} \mathbf{X}\mathbf{X}^\dagger\mathbf{U} = \mathbf{U}\begin{bmatrix} \hat{\mathbf{\Sigma}} \\ 0 \end{bmatrix}\begin{bmatrix} \hat{\mathbf{\Sigma}} & 0 \end{bmatrix} = \mathbf{U}\begin{bmatrix} \hat{\mathbf{\Sigma}}^2 & 0 \\ 0 & 0 \end{bmatrix} \\ \mathbf{X}^\dagger\mathbf{X}\mathbf{V} = \mathbf{V}\begin{bmatrix} \hat{\mathbf{\Sigma}} & 0 \end{bmatrix}\begin{bmatrix} \hat{\mathbf{\Sigma}} \\ 0 \end{bmatrix} = \mathbf{V}\hat{\mathbf{\Sigma}}^2 \end{cases} \tag{2}$$

So the columns of $\mathbf{U}$ are eigenvectors of $\mathbf{X}\mathbf{X}^\dagger$ and the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{X}^\dagger\mathbf{X}$. We choose to arrange the singular values in descending order by magnitude, and thus the columns of $\mathbf{U}$ are hierarchically ordered by how much correlation they capture in the columns of $\mathbf{X}$; $\mathbf{V}$ similarly captures correlation in the rows of $\mathbf{X}$.

## 4 MATRIX APROXIMATION

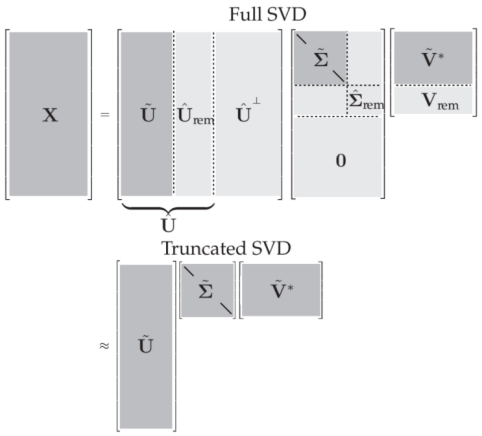
### 4.1 TRUNCATION

Perhaps the most useful and defining property of the SVD is that it provides a hierarchical optimal low-rank approximation to a matrix $\mathbf{X}$.

**Theorem 1 (Eckart-Young)** *The optimal rank-r approximation to* $\mathbf{X}$*, in a least squares sense, is given by the rank r SVD truncation* $\tilde{\mathbf{X}}$*:*

$$\underset{\text{s.t. } \mathrm{rank}(\tilde{\mathbf{X}})=r}{\mathrm{argmin}} \quad ||\mathbf{X} - \tilde{\mathbf{X}}||_F = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\dagger.$$

*where* $||A||_F^2 = \sum_{ij} A_{ij}^2$.




Deciding how many singular values to keep, i.e. where to truncate, is one of the most important and contentious decisions. There are many factors, including specifications on the desired rank of the system, the magnitude of noise, and the distribution of the singular values. Often, one truncates the SVD at a rank r that captures a pre-determined amount of the variance or energy in the original data, such as 90% or 99% truncation. Although crude, this technique is commonly used. Other techniques involve identifying "elbows" or "knees" in the singular value distribution, which may denote the transition from singular values that represent important patterns from those that represent noise. Truncation may be viewed as a hard threshold on singular values, where values larger than a threshold $\tau$ are kept, while remaining singular values are truncated. Recent work by Gavish and Donoho provides an optimal truncation value, or hard threshold, under certain conditions, providing a principled approach to obtaining low-rank matrix approximations using the SVD.

### 4.1.1 Optimal Truncation

Assume that the data matrix $\mathbf{X}$ is the sum of an underlying low-rank, or approximately low-rank, matrix $\mathbf{X}_{true}$ and a noise matrix $\mathbf{X}_{noise}$:

$$\mathbf{X} = \mathbf{X_{true}} + \gamma \mathbf{X_{noise}}$$

The entries of $\mathbf{X}_{noise}$ are assumed to be independent, identically distributed (i.i.d.) Gaussian random variables with zero mean and unit variance. The magnitude of the noise is characterized by $\gamma$. When the noise magnitude $\gamma$ is known, there are closed-form solutions for the optimal hard threshold $\tau$:

1. If $\mathbf{X} \in \mathbb{R}^{n \times n}$ is square, then:

$$\tau = \frac{4}{\sqrt{3}} \sqrt{n} \gamma$$

2. If $\mathbf{X} \in \mathbb{R}^{n \times m}$ is rectangular and $m << n$, then:

$$\tau = \lambda(\beta) \sqrt{n} \gamma$$

$$\lambda(\beta) = \left( 2(\beta + 1) + \frac{8\beta}{(\beta + 1) + (\beta^2 + 14\beta + 1)^{1/2}} \right)^{1/2}$$

where $\beta = m/n$. In case $n << m$ the previous equation holds for $\beta = n/m$. When the noise magnitude $\gamma$ is unknown, which is more typical in real-world applications, then it is possible to estimate the noise magnitude and scale the distribution of singular values by using $\sigma_{med}$, the median singular value. In this case, there is no closed form solution and it must be approximated numerically:

3. For unknown noise $\gamma$, and a rectangular matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, the optimal hard threshold is given by:

$$\tau = \omega(\beta) \sigma_{med}$$

Here, $\omega(\beta) = \lambda(\beta)/\mu_\beta$, where $\mu_\beta$ is the solution to the following problem:

$$\int_{(1-\beta)^2}^{\mu_\beta} \frac{\left[ \left( (1 + \sqrt{\beta})^2 - t \right) \left( t - (1 - \sqrt{\beta})^2 \right) \right]^{1/2}}{2\pi t} dt = \frac{1}{2}$$

## 4.2 Method of Snapshots

If $\mathbf{X}$ is so big that you can not load it in memory you can load two vectors at a time and compute element by element the correlation matrix $\mathbf{X}^\dagger \mathbf{X}$ and then compute its eigenvectors $\mathbf{V}$ and eigenvalues $\hat{\mathbf{\Sigma}}$. Then you can compute $\mathbf{U} = \mathbf{X} \mathbf{V} \hat{\mathbf{\Sigma}}^{-1}$ column by column.

# 5 Pseudo-Inverse, Least-Squares, and Regression

Many physical systems may be represented as a linear system of equations:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

where the constraint matrix $\mathbf{A}$ and vector $\mathbf{b}$ are known, and the vector $\mathbf{x}$ is unknown. If $\mathbf{A}$ is a square, invertible matrix (i.e., $\mathbf{A}$ has nonzero determinant), then there exists a unique solution $\mathbf{x}$ for every $\mathbf{b}$. However, when $\mathbf{A}$ is either singular or rectangular, there may be one, none, or infinitely many solutions, depending on the specific $\mathbf{b}$ and the column and row spaces of $\mathbf{A}$.

In the underdetermined system, where there are fewer equations than unknowns, is likely to have full column rank since it has many more columns than the required for a linearly independent basis. Generically, if a short-fat $\mathbf{A}$ has full column rank, then there are infinitely many solutions $\mathbf{x}$ for every $\mathbf{b}$. The system is called underdetermined because there are not enough values in $\mathbf{b}$ to uniquely determine the higher-dimensional $\mathbf{x}$.

Similarly, consider the overdetermined system, where there are more equations than unknowns. This matrix cannot have a full column rank, and so it is guaranteed that there are vectors $\mathbf{b}$ that have no solution $\mathbf{x}$. In fact, there will only be a solution $\mathbf{x}$ if $\mathbf{b}$ is in the column space of $\mathbf{A}$.

- The column space, $col(\mathbf{A})$, is the span of the columns of $\mathbf{A}$, also known as the range or $Im(\mathbf{A})$. Due that $K^m = Im(\mathbf{A}) \oplus Im(\mathbf{A})^\perp$ and that the orthogonal complement to $col(\mathbf{A})$ is $ker(\mathbf{A}^\dagger)$[1] it follows that
$$K^m = Im(\mathbf{A}) \oplus ker(\mathbf{A}^\dagger)$$
- The row space, $row(\mathbf{A})$, is the span of the rows of $\mathbf{A}$ and it is equal to $Im(\mathbf{A}^\dagger)$. Due that $K^n = ker(\mathbf{A}) \oplus ker(\mathbf{A})^\perp$ and that the orthogonal complement to $ker(\mathbf{A})$ is $Im(\mathbf{A}^\dagger)$it follows that
$$K^n = ker(\mathbf{A}) \oplus Im(\mathbf{A}^\dagger)$$

More precisely, if $\mathbf{b} \in col(\mathbf{A})$ and if $dim(ker(\mathbf{A})) = 0$, then there are infinitely many solutions $\mathbf{x}$. Note that the condition $dim(ker(\mathbf{A})) = 0$ is guaranteed for a short-fat matrix. Similarly, if $\mathbf{b} \notin col(\mathbf{A})$, then there are no solutions, and the system of equations are called inconsistent.

In the overdetermined case when no solution exists, we would often like to find the solution $\mathbf{x}$ that minimizes the sum-squared error $||\mathbf{Ax} - \mathbf{b}||_2^2$, the so-called least-squares solution. Note that the least-squares solution also minimizes $||\mathbf{Ax} - \mathbf{b}||_2$. In the underdetermined case when infinitely many solutions exist, we may like to find the solution $\mathbf{x}$ with minimum norm $||\mathbf{x}||$ so that $\mathbf{Ax} = \mathbf{b}$, the so-called minimum-norm solution. The SVD is the technique of choice for these important optimization problems.

First, if we substitute an exact truncated SVD $\mathbf{A} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\dagger$ in for $\mathbf{A}$, we can "invert" each of the matrices, resulting in the Moore-Penrose left pseudo-inverse $\mathbf{A}^\dagger$ of $\mathbf{A}$:

$$\mathbf{A}^\dagger := \tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}^{-\mathbf{1}}\tilde{\mathbf{U}}^\dagger \Rightarrow \mathbf{A}^\dagger\mathbf{A} = I_{m\times m}$$

This may be used to find both the minimum norm and least-squares solutions to:

$$\mathbf{A}^\dagger\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{b} \Rightarrow \tilde{\mathbf{x}} = \tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}^{-\mathbf{1}}\tilde{\mathbf{U}}^\dagger\mathbf{b}$$

Plugging the solution $\tilde{\mathbf{x}}$ back:

$$\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\dagger\mathbf{b}$$

Note that $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\dagger$ is not necessarily the identity matrix, but is rather a projection onto the column space of $\tilde{\mathbf{U}}$. Therefore, $\tilde{\mathbf{x}}$ will only be an exact solution to when $\mathbf{b}$ is in the column space of $\tilde{\mathbf{U}}$, and therefore in the column space of $\mathbf{A}$.

## 6 Principal Component Analysis

Principal Component Analysis (PCA) is one of the central uses of the SVD, providing a data-driven, hierarchical coordinate system to represent high-dimensional correlated data. This coordinate system involves the correlation matrices described previously. Importantly, PCA pre-processes the data by mean subtraction and setting the variance to unity before performing the SVD. The geometry of the resulting coordinate system is determined by principal components (PCs) that are uncorrelated (orthogonal) to each other, but have maximal correlation with the measurements.

In order to be consistent with PCA literature we swich convention for X, consisting of rows of features instead of arrange them as columns.

[1] $\mathbf{x} \in ker(\mathbf{A}^\dagger) \iff \mathbf{x} \cdot row_i(\mathbf{A}^\dagger) = 0, \forall i \iff \mathbf{x} \cdot col_i(\mathbf{A}) = 0, \forall i \iff \mathbf{x} \in Im(\mathbf{A})^\perp$

### 6.1 COMPUTATION

We now compute the row-wise mean $\tilde{\mathbf{x}}$ (the mean of all rows) and subtract it from $\mathbf{X}$. The mean $\tilde{\mathbf{x}}$ is given by:

$$\tilde{\mathbf{x}}_\mathbf{j} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{\mathbf{ij}}$$

and the mean matrix is

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tilde{\mathbf{x}}$$

Subtracting $\tilde{\mathbf{X}}$ from $\mathbf{X}$ results in the mean-subtracted data $\mathbf{B}$:

$$\mathbf{B} = \mathbf{X} - \tilde{\mathbf{B}}$$

The co-variance matrix of the rows of $\mathbf{B}$ is given by

$$\mathbf{C} = \frac{1}{n-1}\mathbf{B}^*\mathbf{B}$$

the principal component $\mathbf{u_1}$ is given as

$$\mathbf{u_1} = \underset{||\mathbf{u_1}=1||}{\text{argmax}} \quad \mathbf{u_1^*}\mathbf{B}^*\mathbf{B}\mathbf{u_1}$$

which is the eigenvector of $\mathbf{B}^*\mathbf{B}$ corresponding to the largest eigenvalue. Now it is clear that $\mathbf{u_1}$ is the left singular value of $\mathbf{B}$ corresponding to the largest singular value. it is possible to obtain the principal components by computing the eigen-decomposition of $\mathbf{C}$:

$$\mathbf{CV} = \mathbf{VD}$$

which is guarateed to exist, since $\mathbf{C}$ is Hermitian.

## 7 RANDOMIZED SINGULAR VALUE DECOMPOSITION

The accurate and efficient decomposition of large data matrices is one of the cornerstones of modern computational mathematics and data science. Recently, it has been shown that if a matrix $\mathbf{X}$ has low-rank structure, then there are extremely efficient matrix decomposition algorithms based on the theory of random sampling. These so-called randomized numerical methods have the potential to provide accurate matrix decompositions at a fraction of the cost of deterministic methods. Most randomized matrix decompositions can be broken into a few common steps,

Step 0: Identify a target rank, $r < m$.

Step 1: Using random projections $\mathbf{P}$ to sample the column space, find a matrix $\mathbf{Q}$ whose columns approximate the column space of $\mathbf{X}$, so that $\mathbf{X} \sim \mathbf{QQ}^\star\mathbf{X}$ .

Step 2: Project $\mathbf{X}$ onto the $\mathbf{Q}$ subspace, $\mathbf{Y} = \mathbf{QX}$, and compute the matrix decomposition on $\mathbf{Y}$.

Step 3: Reconstruct high dimensional modes $\mathbf{U} = \mathbf{QUY}$ using $\mathbf{Q}$ and the modes computed from Y.