The Singular Value Decomposition

Carlo Tomasi

September 16, 2017

Section 1 defines the concepts of orthogonality and projection for general $m \times n$ matrices. The Sections thereafter use these concepts to introduce the Singular Value Decomposition (SVD) of a matrix and principal component analysis. When not given in the main text, proofs are in Appendix A.

1 Orthogonal Matrices

Let $\mathcal{S}$ be an n-dimensional subspace of $\mathbf{R}^m$ (so that we necessarily have $n \leq m$), and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be an orthonormal basis for $\mathcal{S}$. Consider a point P in $\mathcal{S}$. If the coordinates of P in $\mathbf{R}^m$ are collected in an m-dimensional vector

$$\mathbf{p} = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} ,$$

and since P is in $\mathcal{S}$, it must be possible to write $\mathbf{p}$ as a linear combination of the $\mathbf{v}_j$s. In other words, there must exist coefficients

$$\mathbf{q} = \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}$$

such that

$$\mathbf{p} = q_1\mathbf{v}_1 + \ldots + q_n\mathbf{v}_n = V\mathbf{q}$$

where

$$V = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix}$$

is an $m \times n$ matrix that collects the basis for $\mathcal{S}$ as its columns. Then for any $i = 1, \ldots, n$ we have

$$\mathbf{v}_i^T\mathbf{p} = \mathbf{v}_i^T \sum_{j=1}^{n} q_j\mathbf{v}_j = \sum_{j=1}^{n} q_j\mathbf{v}_i^T\mathbf{v}_j = q_i ,$$

since the $\mathbf{v}_j$ are orthonormal. This is important, and may need emphasis:

If

$$\mathbf{p} = \sum_{j=1}^{n} q_j\mathbf{v}_j$$

and the vectors of the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are orthonormal, then the coefficients q_j are the signed magnitudes of the projections of $\mathbf{p}$ onto the basis vectors:

$$q_j = \mathbf{v}_j^T\mathbf{p} . \tag{1}$$

In matrix form,

$$\mathbf{q} = V^T \mathbf{p} \ . \tag{2}$$

Also, we can collect the n^2 equations

$$\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

into the following matrix equation:

$$V^T V = I \tag{3}$$

where I is the $n \times n$ identity matrix. A matrix V that satisfies equation (3) is said to be *orthogonal*. Thus, a matrix is orthogonal if its columns are orthonormal. Since the *left inverse* of a matrix V is defined as the matrix L such that

$$LV = I \ , \tag{4}$$

comparison with equation (3) shows that the left inverse of an orthogonal matrix V exists, and is equal to the transpose of V.

Of course, this argument requires V to be full rank, so that the solution L to equation (4) is unique. However, V is certainly full rank, because it is made of orthonormal columns.

Notice that $VR = I$ cannot possibly have a solution when $m > n$, because the $m \times m$ identity matrix has m linearly independent [1] columns, while the columns of VR are linear combinations of the n columns of V, so VR can have at most n linearly independent columns.

Of course, this result is still valid when V is $m \times m$ and has orthonormal columns, since equation (3) still holds. However, for square, full-rank matrices ($r = m = n$), the distinction between left and right inverse vanishes. Since the matrix VV^T contains the inner products between the *rows* of V (just as $V^T V$ is formed by the inner products of its *columns*), the argument above shows that the rows of a *square* orthogonal matrix are orthonormal as well. We can summarize this discussion as follows:

Theorem 1.1. *The left inverse of an orthogonal $m \times n$ matrix V with $m \geq n$ exists and is equal to the transpose of V:*

$$V^T V = I \ .$$

In particular, if $m = n$, the matrix $V^{-1} = V^T$ is also the right inverse of V:

$$V \text{ square} \quad \Rightarrow \quad V^{-1} V = V^T V = V V^{-1} = V V^T = I \ .$$

Sometimes, when $m = n$, the geometric interpretation of equation (2) causes confusion, because two interpretations of it are possible. In the interpretation given above, the point P remains the same, and the underlying reference frame is changed from the elementary vectors $\mathbf{e}_j$ (that is, from the columns of I) to the vectors $\mathbf{v}_j$ (that is, to the columns of V). Alternatively, equation (2) can be seen as a transformation, in a fixed reference system, of point P with coordinates $\mathbf{p}$ into a different point Q with coordinates $\mathbf{q}$. This, however, is relativity, and should not be surprising: If you spin clockwise on your feet, or if you stand still and the whole universe spins counterclockwise around you, the result is the same.[2]

Consistently with either of these geometric interpretations, we have the following result:

[1] Nay, orthonormal.

[2] At least geometrically. One solution may be more efficient than the other in other ways.

Theorem 1.2. *The norm of a vector* $\mathbf{x}$ *is not changed by multiplication by an orthogonal matrix* V*:*

$$\|V\mathbf{x}\| = \|\mathbf{x}\| \ .$$

The proof is a one-liner, so it is included here:

$$\|V\mathbf{x}\|^2 = \mathbf{x}^T V^T V\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2 \ .$$

We conclude this section with an obvious but useful consequence of orthogonality. First, define the *projection* $\mathbf{p}$ of a point $\mathbf{b} \in \mathbf{R}^n$ *onto a subspace* C as the point in C that is closest to $\mathbf{b}$. The following theorem shows how to project a point onto the range of an orthogonal matrix, and how the point and its projection relate to each other.

Theorem 1.3. *Let* U *be an orthogonal matrix. Then the matrix* UU^T *projects any vector* $\mathbf{b}$ *onto* range(U)*. Furthermore, the difference vector between* $\mathbf{b}$ *and its projection* $\mathbf{p}$ *onto* range(U) *is orthogonal to* range(U)*:*

$$U^T(\mathbf{b} - \mathbf{p}) = \mathbf{0} \ .$$

2 The Singular Value Decomposition

Here is the main intuition captured by the Singular Value Decomposition (SVD) of a matrix:

> An $m \times n$ matrix A of rank r maps the r-dimensional unit hypersphere in rowspace(A) into an r-dimensional hyperellipse in range(A).

Thus, a hypersphere is stretched or compressed into a hyperellipse, which is a quadratic hypersurface that generalizes the two-dimensional notion of ellipse to an arbitrary number of dimensions. In three dimensions, the hyperellipse is an ellipsoid, in one dimension it is a pair of points. In all cases, the hyperellipse in question is centered at the origin.

For instance, the rank-2 matrix

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{3} & \sqrt{3} \\ -3 & 3 \\ 1 & 1 \end{bmatrix} \tag{5}$$

transforms the unit circle on the plane into an ellipse embedded in three-dimensional space. Figure 1 shows the map

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

Two diametrically opposite points on the unit circle are mapped into the two endpoints of the major axis of the ellipse, and two other diametrically opposite points on the unit circle are mapped into the two endpoints of the minor axis of the ellipse. The lines through these two pairs of points on the unit circle are always orthogonal. This result can be generalized to any $m \times n$ matrix.

Simple and fundamental as this geometric fact may be, its proof by geometric means is cumbersome. It is, on the other hand, a straightforward consequence of the following fundamental theorem, which states the existence of the SVD.

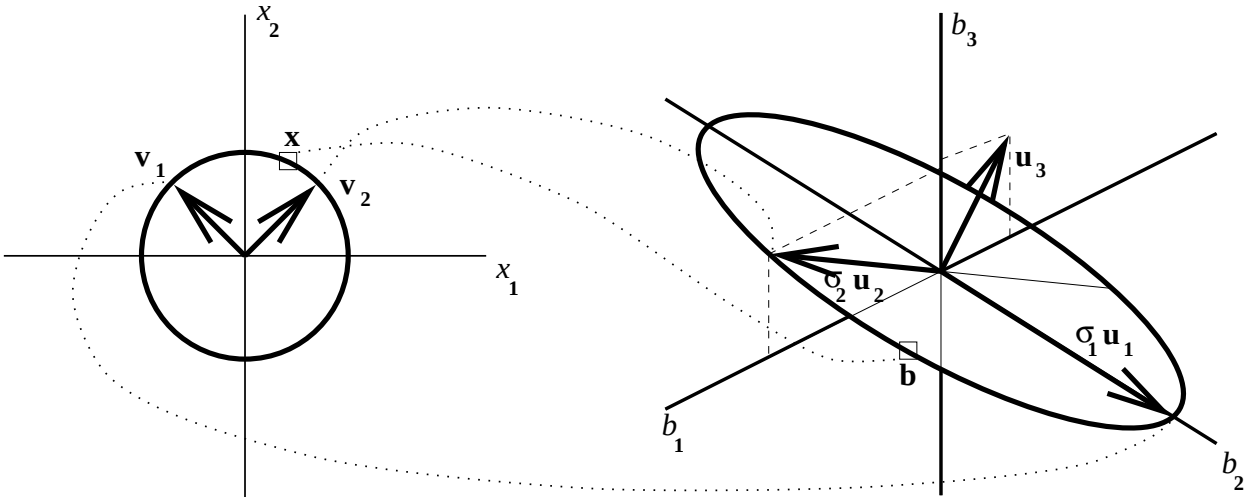

Figure 1: The matrix in equation (5) maps a circle on the plane into an ellipse in space. The two small boxes are corresponding points.

Theorem 2.1. *If A is a real $m \times n$ matrix then there exist orthogonal matrices*

$$\begin{aligned} U &= \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{array} \right] \in \mathcal{R}^{m \times m} \\ V &= \left[\begin{array}{ccc} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{array} \right] \in \mathcal{R}^{n \times n} \end{aligned}$$

such that

$$U^T A V = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathcal{R}^{m \times n}$$

where $p = \min(m, n)$ and $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$. Equivalently,

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

The columns of V are the *right singular vectors* of A, and those of U are its *left singular vectors*. The diagonal entries of Σ are the *singular values* of A. The ratio

$$\kappa(A) = \sigma_1 / \sigma_p \tag{6}$$

is the *condition number* of A, and is possibly infinite.

The singular value decomposition is "almost unique". There are two sources of ambiguity. The first is in the orientation of the singular vectors. One can flip any right singular vector, provided that the corresponding left singular vector is flipped as well, and still obtain a valid SVD. Singular vectors must be flipped in pairs (a left vector and its corresponding right vector) because the singular values are required to be nonnegative. This is a trivial ambiguity. If desired, it can be removed by imposing, for instance, that the first nonzero entry of every left singular value be positive.

The second source of ambiguity is deeper. If the matrix A maps a hypersphere into another hypersphere, the axes of the latter are not defined. For instance, the identity matrix has an infinity of SVDs, all of the form

$$I = U I U^T$$

where U is any orthogonal matrix of suitable size. More generally, whenever two or more singular values coincide, the subspaces identified by the corresponding left and right singular vectors are unique, but any orthonormal basis can be chosen within, say, the right subspace and yield, together with the corresponding left singular vectors, a valid SVD. Except for these ambiguities, the SVD is unique.

Even in the general case, the singular values of a matrix A are the lengths of the semi-axes of the hyperellipse E defined by

$$E = \{A\mathbf{x} : \|\mathbf{x}\| = 1\} .$$

The SVD reveals a great deal about the structure of a matrix. If we define r by

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = 0 ,$$

that is, if σ_r is the smallest nonzero singular value of A, then

$$\begin{aligned} \text{rank}(A) &= r \\ \text{null}(A) &= \text{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\} \\ \text{range}(A) &= \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\} . \end{aligned}$$

The sizes of the matrices in the SVD are as follows: U is $m \times m$, Σ is $m \times n$, and V is $n \times n$. Thus, Σ has the same shape and size as A, while U and V are square. However, if $m > n$, the bottom $(m-n) \times n$ block of Σ is zero, so that the last $m - n$ columns of U are multiplied by zero. Similarly, if $m < n$, the rightmost $m \times (n - m)$ block of Σ is zero, and this multiplies the last $n - m$ rows of V. This suggests a "small," equivalent version of the SVD. If $p = \min(m, n)$, we can define $U_p = U(:, 1 : p)$, $\Sigma_p = \Sigma(1 : p, 1 : p)$, and $V_p = V(:, 1 : p)$, and write

$$A = U_p \Sigma_p V_p^T$$

where U_p is $m \times p$, Σ_p is $p \times p$, and V_p is $n \times p$.

Moreover, if $p - r$ singular values are zero, we can let $U_r = U(:, 1 : r)$, $\Sigma_r = \Sigma(1 : r, 1 : r)$, and $V_r = V(:, 1 : r)$, then we have

$$A = U_r \Sigma_r V_r^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T ,$$

which is an even smaller, *minimal*, SVD.

Finally, both the 2-norm and the Frobenius norm

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

and

$$\|A\|_2 = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \tag{7}$$

are neatly characterized in terms of the SVD:

$$\begin{aligned} \|A\|_F^2 &= \sigma_1^2 + \ldots + \sigma_p^2 \\ \|A\|_2 &= \sigma_1 . \end{aligned}$$

In the next section we introduce a first fundamental application of the SVD.

3 Principal Component Analysis

Principal Component Analysis (PCA) is a lossy data compression technique. Given a set of vectors in a space $\mathbb{R}^m$ with large m, it is often the case that most of the variation in the set occurs along a number k of dimensions that is much smaller than m. PCA finds an orthogonal basis for the smaller, k-dimensional space and projects the data down to it, so that subsequent data processing is more efficient. In machine learning, this compression often also leads to better generalization.

Data variation is measured by empirical covariance, which is in turn an estimate of the statistical covariance of a probability distribution that is assumed to have generated the data. This section recalls the definitions of statistical and empirical covariance, introduces the PCA, and states its main properties, which are proven in the Appendix.

The (statistical) *covariance matrix* of a random vector $\mathbf{a} \in \mathbb{R}^m$ is defined as the $m \times m$ matrix

$$\Sigma_{\mathbf{a}} = \mathbb{E}[(\mathbf{a} - \mathbf{m_a})(\mathbf{a} - \mathbf{m_a})^T] \quad \text{where} \quad \mathbf{m_a} = \mathbb{E}[\mathbf{a}] \ .$$

This matrix describes the spread of the vector around its mean $\mathbf{m_a}$. If the vector is Gaussian, then the ellipsoids with equation

$$(\mathbf{a} - \mathbf{m_a})^T \Sigma_{\mathbf{a}}^{-1} (\mathbf{a} - \mathbf{m_a}) = c$$

are the loci of constant probability density, and

$$\mathbb{P}[(\mathbf{a} - \mathbf{m_a})^T \Sigma_{\mathbf{a}}^{-1} (\mathbf{a} - \mathbf{m_a}) \leq c] = F_{\chi^2_m}(c) \ ,$$

where the right-hand side is the cumulative distribution function of the chi-square random variable with m degrees of freedom. Approximately half of the probability mass of $\mathbf{a}$ is in the ellipsoid for $c = m$, and about 90 percent of the mass is in the ellipsoid for $c = m + 2\sqrt{m}$ (for large m). For $m = 1$, the scalar $\Sigma_a = \sigma_a^2$ is the variance of the random variable a.

The *empirical covariance matrix* for a set of n independent samples $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ from a random variable $\mathbf{a}$ is an unbiased estimate of the covariance of $\mathbf{a}$ and is defined as follows:

$$Q(A) = \frac{1}{n-1} A_c A_c^T \quad \text{where} \quad A_c = A - \boldsymbol{\mu}(A) \mathbf{1}_n^T$$

and $\mathbf{1}_n$ is a column vector of n ones. The vector

$$\boldsymbol{\mu}(A) = \frac{1}{n} A \mathbf{1}_n$$

is an unbiased estimate of the mean of $\mathbf{a}$, so the columns of A_c can be viewed as the result of *centering* the columns of A around their mean.

Given any $m \times n$ matrix $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ and integer $k \leq \min(m, n)$, let

$$A_c = U_c \Sigma_c V_c^T \quad \text{with} \quad \Sigma_c = \operatorname{diag}[\sigma_1, \ldots, \sigma_{\min(m,n)}]$$

be the SVD of the centered data matrix A_c and define the $m \times k$ orthogonal matrix

$$U = U_c(:, 1{:}k) \ .$$

The process of computing the $k \times n$ matrix

$$B = U^T A_c \tag{8}$$

is called the *Principal Component Analysis (PCA)* of A. Algorithm 1 summarizes the computation. Sometimes, the matrix B itself is called the PCA of A.

Algorithm 1 Principal Component Analysis

Input: An $m \times n$ matrix A (the columns of A are the data points) and an integer $k \leq \min(m, n)$

$\boldsymbol{\mu} \leftarrow A\mathbf{1}_n/n$

$A_c \leftarrow A - \boldsymbol{\mu}\mathbf{1}_n^T$

$[U, S, \sim] \leftarrow \mathrm{svd}(A_c)$ ▷ The "small" SVD suffices, and the matrix V of right singular vectors is unused

$U \leftarrow U(:, 1{:}k)$

$\mathbf{s} \leftarrow \mathrm{diag}(S)/\sqrt{n-1}$ ▷ These are standard deviations, not variances

$B \leftarrow U^T A_c$ ▷ The columns of B are the transformed data

return U, $\boldsymbol{\mu}$, B, $\mathbf{s}$

Output: An $m \times k$ matrix U with $U^T U = I_k$, the m-dimensional centroid $\boldsymbol{\mu}$ of the data in A, a $k \times n$ matrix B, and a vector $\mathbf{s}$ of $\min(m, n)$ standard deviations

Theorem 3.1. *Let A be an $m \times n$ matrix of n data points in $\mathbb{R}^m$ whose centered matrix $A_c = A - \boldsymbol{\mu}(A)\mathbf{1}_n^T$ has singular values $\sigma_1, \ldots, \sigma_{\min(m,n)}$. Also, let k be an integer no greater than $\min(m, n)$. The columns of the $k \times n$ matrix $B = U^T A_c$ computed by PCA of A enjoy the following properties:*

- *They live in a space $\mathbb{R}^k$ with dimensionality no greater than m,*

$$k \leq m \ .$$

- *They are uncorrelated, with empirical covariance matrix*

$$Q(B) = \frac{1}{n-1}\mathrm{diag}[\sigma_1^2, \ldots, \sigma_k^2] \quad \textit{with} \quad \sigma_1^2 \geq \ldots \geq \sigma_k^2 \ .$$

- *They capture the dimensions of greatest variance in A, in the sense that*

$$\|A - UB\|_2 = \sup_{\|\mathbf{x}\| \neq 0} \frac{\|(A - UB)\mathbf{x}\|}{\|\mathbf{x}\|} = \sigma_{k+1} \leq \sigma_k \ .$$

In geometric terms, if we approximate the distribution of the columns of A with an m-dimensional ellipsoid $\mathcal{E}(A)$ centered at $\boldsymbol{\mu}(A)$ and with covariance $Q(A)$, then the k-dimensional ellipsoid $\mathcal{E}(B)$ that approximates the columns of B is centered at the origin. Its k axes are equal in length to the k longest axes of $\mathcal{E}(A)$ and are aligned with the axes of the new reference system (because $Q(B)$ is diagonal).

A diagonal covariance matrix also means that the components (*i.e.*, the coordinates) of the transformed data in B are mutually uncorrelated. As a consequence, variances along the diagonal add. The variance in B is $\sigma_1^2 + \ldots + \sigma_k^2$, while the variance in A is $\sigma_1^2 + \ldots + \sigma_n^2$. Since the singular values are listed in non-increasing order along the diagonal of $Q(B)$, the transformation performed by the matrix U preserves the greatest possible fraction of the variance of the data in A among all linear projections, and that fraction is

$$\frac{\sigma_1^2 + \ldots + \sigma_k^2}{\sigma_1^2 + \ldots + \sigma_n^2} \ .$$

CAVEAT: Preserving variance is often useful, but is not always what is needed. For instance, most of the variance of the data in Figure 2 is in the horizontal direction. However, if plusses and minuses represent two different classes in a classification problem, then the component that matters most for classification is

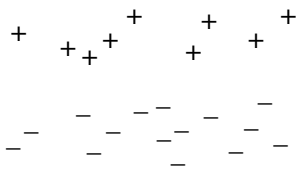

Figure 2: PCA would project this data onto an approximately horizontal line, obliterating most class-discriminative information in the process.

the vertical one. In this example, taking a $k = 1$-dimensional PCA of this $m = 2$-dimensional data set would project all points onto a horizontal line, irreversibly mixing plusses and minuses. So while PCA can help by reducing the size of the feature space, it can also hurt by removing information that is relevant for classification. Because of this, the value of k should be chosen with care, and dimensionality reduction should not be too aggressive if the data is used for classification.

A Proofs

Theorem 1.3

Let U be an orthogonal matrix. Then the matrix UU^T projects any vector $\mathbf{b}$ onto range(U). *Furthermore, the difference vector between $\mathbf{b}$ and its projection $\mathbf{p}$ onto* range(U) *is orthogonal to* range(U)*:*

$$U^T(\mathbf{b} - \mathbf{p}) = \mathbf{0} \ .$$

Proof. A point $\mathbf{p}$ in range(U) is a linear combination of the columns of U:

$$\mathbf{p} = U\mathbf{x}$$

where $\mathbf{x}$ is the vector of coefficients (as many coefficients as there are columns in U). The squared distance between $\mathbf{b}$ and $\mathbf{p}$ is

$$\|\mathbf{b} - \mathbf{p}\|^2 = (\mathbf{b} - \mathbf{p})^T(\mathbf{b} - \mathbf{p}) = \mathbf{b}^T\mathbf{b} + \mathbf{p}^T\mathbf{p} - 2\mathbf{b}^T\mathbf{p} = \mathbf{b}^T\mathbf{b} + \mathbf{x}^TU^TU\mathbf{x} - 2\mathbf{b}^TU\mathbf{x} \ .$$

Because of orthogonality, U^TU is the identity matrix, so

$$\|\mathbf{b} - \mathbf{p}\|^2 = \mathbf{b}^T\mathbf{b} + \mathbf{x}^T\mathbf{x} - 2\mathbf{b}^TU\mathbf{x} \ .$$

The derivative of this squared distance with respect to $\mathbf{x}$ is the vector

$$2\mathbf{x} - 2U^T\mathbf{b}$$

which is zero iff

$$\mathbf{x} = U^T\mathbf{b} \ ,$$

that is, when

$$\mathbf{p} = U\mathbf{x} = UU^T\mathbf{b}$$

as promised.

For this value of $\mathbf{p}$ the difference vector $\mathbf{b} - \mathbf{p}$ is orthogonal to range(U), in the sense that

$$U^T(\mathbf{b} - \mathbf{p}) = U^T(\mathbf{b} - UU^T\mathbf{b}) = U^T\mathbf{b} - U^T\mathbf{b} = \mathbf{0} \ .$$

Theorem 2.1

If A is a real $m \times n$ matrix then there exist orthogonal matrices

$$\begin{aligned} U &= \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{array} \right] \in \mathcal{R}^{m \times m} \\ V &= \left[\begin{array}{ccc} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{array} \right] \in \mathcal{R}^{n \times n} \end{aligned}$$

such that

$$U^T A V = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathcal{R}^{m \times n}$$

where $p = \min(m, n)$ and $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$. Equivalently,

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

Proof. Let $\mathbf{x}$ and $\mathbf{y}$ be unit vectors in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and consider the bilinear form

$$z = \mathbf{y}^T A \mathbf{x} \ .$$

The set

$$\mathcal{S} \ = \ \{\mathbf{x}, \, \mathbf{y} \mid \mathbf{x} \in \mathbf{R}^n, \ \mathbf{y} \in \mathbf{R}^m, \ \|\mathbf{x}\| = \|\mathbf{y}\| = 1\}$$

is compact, so that the scalar function $z(\mathbf{x}, \mathbf{y})$ must achieve a maximum value on $\mathcal{S}$, possibly at more than one point [3]. Let $\mathbf{u}_1$, $\mathbf{v}_1$ be two unit vectors in $\mathbf{R}^m$ and $\mathbf{R}^n$ respectively where this maximum is achieved, and let σ_1 be the corresponding value of z:

$$\max_{\|\mathbf{x}\| = \|\mathbf{y}\| = 1} \mathbf{y}^T A \mathbf{x} = \mathbf{u}_1^T A \mathbf{v}_1 = \sigma_1 \ .$$

It is easy to see that $\mathbf{u}_1$ is parallel to the vector $A\mathbf{v}_1$. If this were not the case, their inner product $\mathbf{u}_1^T A \mathbf{v}_1$ could be increased by rotating $\mathbf{u}_1$ towards the direction of $A\mathbf{v}_1$, thereby contradicting the fact that $\mathbf{u}_1^T A \mathbf{v}_1$ is a maximum. Similarly, by noticing that

$$\mathbf{u}_1^T A \mathbf{v}_1 = \mathbf{v}_1^T A^T \mathbf{u}_1$$

and repeating the argument above, we see that $\mathbf{v}_1$ is parallel to $A^T \mathbf{u}_1$.

The vectors $\mathbf{u}_1$ and $\mathbf{v}_1$ can be extended into orthonormal bases for $\mathbf{R}^m$ and $\mathbf{R}^n$, respectively. Collect these orthonormal basis vectors into orthogonal matrices U_1 and V_1. Then

$$U_1^T A V_1 = S_1 = \left[\begin{array}{cc} \sigma_1 & \mathbf{0}^T \\ \mathbf{0} & A_1 \end{array} \right] \ .$$

In fact, the first column of AV_1 is $A\mathbf{v}_1 = \sigma_1 \mathbf{u}_1$, so the first entry of $U_1^T A V_1$ is $\mathbf{u}_1^T \sigma_1 \mathbf{u}_1 = \sigma_1$, and its other entries are $\mathbf{u}_j^T A \mathbf{v}_1 = 0$ because $A\mathbf{v}_1$ is parallel to $\mathbf{u}_1$ and therefore orthogonal, by construction, to $\mathbf{u}_2, \ldots, \mathbf{u}_m$. A similar argument shows that the entries after the first in the first row of S_1 are zero: the row vector $\mathbf{u}_1^T A$ is parallel to $\mathbf{v}_1^T$, and therefore orthogonal to $\mathbf{v}_2, \ldots, \mathbf{v}_n$, so that $\mathbf{u}_1^T A \mathbf{v}_2 = \ldots = \mathbf{u}_1^T A \mathbf{v}_n = 0$.

The matrix A_1 has one fewer row and column than A. We can repeat the same construction on A_1 and write

$$U_2^T A_1 V_2 = S_2 = \left[\begin{array}{cc} \sigma_2 & \mathbf{0}^T \\ \mathbf{0} & A_2 \end{array} \right]$$

[3] Actually, at least at two points: if $\mathbf{u}_1^T A \mathbf{v}_1$ is a maximum, so is $(-\mathbf{u}_1)^T A (-\mathbf{v}_1)$.

so that

$$\begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & U_2^T \end{bmatrix} U_1^T A V_1 \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & V_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & \mathbf{0}^T \\ 0 & \sigma_2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & A_2 \end{bmatrix} .$$

This procedure can be repeated until A_k vanishes (zero rows or zero columns) to obtain

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

where U^T and V are orthogonal matrices obtained by multiplying together all the orthogonal matrices used in the procedure, and

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) .$$

Since matrices U and V are orthogonal, we can premultiply the matrix product in the theorem by U and postmultiply it by V^T to obtain

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

which is the desired result.

It only remains to show that the elements on the diagonal of Σ are nonnegative and arranged in non-increasing order. To see that $\sigma_1 \geq \ldots \geq \sigma_p$ (where $p = \min(m, n)$), we can observe that the successive maximization problems that yield $\sigma_1, \ldots, \sigma_p$ are performed on a sequence of sets each of which contains the next. To show this, we just need to show that $\sigma_2 \leq \sigma_1$, and induction will do the rest. We have

$$\begin{aligned} \sigma_2 &= \max_{\|\hat{\mathbf{x}}\|=\|\hat{\mathbf{y}}\|=1} \hat{\mathbf{y}}^T A_1 \hat{\mathbf{x}} = \max_{\|\hat{\mathbf{x}}\|=\|\hat{\mathbf{y}}\|=1} \begin{bmatrix} 0 & \hat{\mathbf{y}} \end{bmatrix}^T S_1 \begin{bmatrix} 0 \\ \hat{\mathbf{x}} \end{bmatrix} \\ &= \max_{\|\hat{\mathbf{x}}\|=\|\hat{\mathbf{y}}\|=1} \begin{bmatrix} 0 & \hat{\mathbf{y}} \end{bmatrix}^T U_1^T A V_1 \begin{bmatrix} 0 \\ \hat{\mathbf{x}} \end{bmatrix} = \max_{\substack{\|\mathbf{x}\| = \|\mathbf{y}\| = 1 \\ \mathbf{x}^T \mathbf{v}_1 = \mathbf{y}^T \mathbf{u}_1 = 0}} \mathbf{y}^T A \mathbf{x} \leq \sigma_1 . \end{aligned}$$

To explain the last equality above, consider the vectors

$$\mathbf{x} = V_1 \begin{bmatrix} 0 \\ \hat{\mathbf{x}} \end{bmatrix} \quad \text{and} \quad \mathbf{y} = U_1 \begin{bmatrix} 0 \\ \hat{\mathbf{y}} \end{bmatrix} .$$

The vector $\mathbf{x}$ is equal to the unit vector $[0\ \hat{\mathbf{x}}]^T$ transformed by the orthogonal matrix V_1, and is therefore itself a unit vector. In addition, it is a linear combination of $\mathbf{v}_2, \ldots, \mathbf{v}_n$, and is therefore orthogonal to $\mathbf{v}_1$. A similar argument shows that $\mathbf{y}$ is a unit vector orthogonal to $\mathbf{u}_1$. Because $\mathbf{x}$ and $\mathbf{y}$ thus defined belong to subsets (actually sub-spheres) of the unit spheres in $\mathbf{R}^n$ and $\mathbf{R}^m$, we conclude that $\sigma_2 \leq \sigma_1$.

The σ_i are nonnegative because all these maximizations are performed on unit hyper-spheres. The σ_is are maxima of the function $z(\mathbf{x}, \mathbf{y})$ which always assumes both positive and negative values on any hyper-sphere: If $z(\mathbf{x}, \mathbf{y})$ is negative, then $z(-\mathbf{x}, \mathbf{y})$ is positive, and if $\mathbf{x}$ is on a hyper-sphere, so is $-\mathbf{x}$.

Theorem 3.1

Let A be an $m \times n$ matrix of n data points in $\mathbb{R}^m$ whose centered matrix $A_c = A - \boldsymbol{\mu}(A)\mathbf{1}_n^T$ has singular values $\sigma_1, \ldots, \sigma_{\min(m,n)}$. Also, let k be an integer no greater than $\min(m,n)$. The columns of the $k \times n$ matrix $B = U^T A_c$ computed by PCA of A enjoy the following properties:

- *They live in a space $\mathbb{R}^k$ with dimensionality no greater than m,*

$$k \leq m \ .$$

- *They are uncorrelated, with empirical covariance matrix*

$$Q(B) = \frac{1}{n-1}\mathrm{diag}[\sigma_1^2, \ldots, \sigma_k^2] \quad \textit{with} \quad \sigma_1^2 \geq \ldots \geq \sigma_k^2 \ .$$

- *They capture the dimensions of greatest variance in A, in the sense that*

$$\|A - UB\|_2 = \sup_{\|\mathbf{x}\| \neq 0} \frac{\|(A-UB)\mathbf{x}\|}{\|\mathbf{x}\|} = \sigma_{k+1} \leq \sigma_k \ .$$

Proof. The first property of B is immediate. For the second property, we have

$$\begin{aligned} (n-1)Q(B) &= BB^T = U^T A_c A_c^T U = U^T U_c \Sigma_c V_c^T (U_c \Sigma_c V_c^T)^T U \\ &= U^T U_c \Sigma_c V_c^T V_c \Sigma_c U_c^T U = U^T U_c \Sigma_c^2 U_c^T U \end{aligned}$$

where

$$U_c^T U = U_c^T U_c(:, 1{:}k) = [I_k \,|\, 0_{k\times(n-k)}]$$

so that

$$(n-1)Q(B) = \Sigma_c^2(1{:}k, 1{:}k) \ .$$

To prove the third property of B, let us first understand the meaning of the *residual matrix*

$$R = A - UB = A - UU^T A = (I_m - UU^T)A = LL^T A$$

where I_m is the $m \times m$ identity matrix and the columns of

$$L = U_c(:, (k+1){:}n)$$

span the left null space of A (see theorem 1.3). The transformation (8) can be written as

$$\mathbf{b}_j = U^T \mathbf{a}_j \quad \text{for} \quad j = 1, \ldots, n \ ,$$

so equation (2) implies that the entries of $\mathbf{b}_j$ are the coefficients of $\mathbf{a}_j$ in the orthonormal basis spanned by the columns of U. Thus, B contains the coefficients of the projection of the columns of A onto $\mathcal{R}_T$, the range of U, and the part of A that B fails to capture is the projection of the columns of A onto $\mathcal{L}_T$, the left null space of U.

Since LL^T projects onto $\mathcal{R}_T$, the greatest singular value of R is

$$\sup_{\|\mathbf{x}\| \neq 0} \frac{\|(A-UB)\mathbf{x}\|}{\|\mathbf{x}\|} = \sigma_{k+1} \leq \sigma_k$$

where the last inequality follows from the ordering property of the singular values. Thus, the 2-norm of the part of the data in A that B fails to capture is at most equal to the standard deviation σ_k along the direction of least variance in B.