ELE 535
Machine Learning and Pattern Recognition [1]
Handout #2: SVD and PCA

Peter J. Ramadge

Fall 2016, version 2.2

Chapter 5

Singular Value Decomposition

5.1 Introduction

This chapter reviews the Singular Value Decomposition (SVD) of a rectangular matrix in both its compact and full form. This very useful matrix factorization extends the idea of an eigendecomposition of a square matrix to non-square matrices. It is useful in general, but has specific application in data analysis, dimensionality reduction (PCA), low rank matrix approximation, and some forms of regression.

5.2 Preliminaries

5.2.1 Range and Null Space

Let $A \in \mathbb{R}^{m\times n}$. The *range* of A, denoted by $\mathcal{R}(A)$, is the subspace of $\mathbb{R}^m$ defined by $\mathcal{R}(A) = \{y\colon y = Ax, \text{ some } x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$. So the range of A is the set of all vectors that can be formed as a linear combination of the columns of A. The *nullspace* of A, denoted by $\mathcal{N}(A)$, is the subspace of $\mathbb{R}^n$ defined by $\mathcal{N}(A) = \{x\colon Ax = \mathbf{0}\}$. This is the set of all vectors that are mapped to the zero vector in $\mathbb{R}^m$ by A.

The following fundamental result from linear algebra will be very useful.

Theorem 5.2.1. Let $A \in \mathbb{R}^{m\times n}$ have nullspace $\mathcal{N}(A)$ and range $\mathcal{R}(A)$. Then $\mathcal{N}(A)^\perp = \mathcal{R}(A^T)$.

Proof. Let $x \in \mathcal{N}(A)$. Then $Ax = \mathbf{0}$ and $x^TA^T = \mathbf{0}^T$. So for every $y \in \mathbb{R}^m$, $x^T(A^Ty) = 0$. Thus $x \in \mathcal{R}(A^T)^\perp$. This shows that $\mathcal{N}(A) \subseteq \mathcal{R}(A^T)^\perp$. Now for all subspaces: (a) $(\mathcal{U}^\perp)^\perp = \mathcal{U}$, and (b) $\mathcal{U} \subseteq \mathcal{V}$ implies $\mathcal{V}^\perp \subseteq \mathcal{U}^\perp$. Applying these properties yields $\mathcal{R}(A^T) \subseteq \mathcal{N}(A)^\perp$.

Conversely, suppose $x \in \mathcal{R}(A^T)^\perp$. Then for all $y \in \mathbb{R}^m$, $x^TA^Ty = 0$. Hence for all $y \in \mathbb{R}^m$, $y^TAx = 0$. This implies $Ax = 0$ and hence that $x \in \mathcal{N}(A)$. Thus $\mathcal{R}(A^T)^\perp \subseteq \mathcal{N}(A)$ and $\mathcal{N}(A)^\perp \subseteq \mathcal{R}(A^T)$. We have shown $\mathcal{R}(A^T) \subseteq \mathcal{N}(A)^\perp$ and $\mathcal{N}(A)^\perp \subseteq \mathcal{R}(A^T)$. Thus $\mathcal{N}(A)^\perp = \mathcal{R}(A^T)$. □

5.2.2 Matrix rank

The *rank* of A is the dimension of the range of A. This equals the number of linearly independent columns in A. The rank of A also equals the number of linearly independent rows of A. Thus $r \leq \min(m, n)$. The matrix A is said to *full rank* if $r = \min(m, n)$.

An $m \times n$ *rank one matrix* has the form yx^T where $y \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ are both nonzero. Notice that for all $w \in \mathbb{R}^n$, $(yx^T)w = y(x^Tw)$ is a scalar multiple of y. Moreover, by suitable choice of w we can make this scalar any real value. So $\mathcal{R}(yx^T) = \text{span}(y)$ and the rank of yx^T is indeed one.

5.2.3 Eigenvalues and Eigenvectors

The *eigenvectors* of a square matrix $A \in \mathbb{C}^{n\times n}$ are the *nonzero* vectors $x \in \mathbb{C}^n$ such that for some $\lambda \in \mathbb{C}$, called the *eigenvalue* corresponding to x, $Ax = \lambda x$. The set of eigenvalues of A is referred to as the *spectrum* of A, and is denoted by $\sigma(A)$. Note that an eigenvector for the eigenvalue λ is not unique; if x is an eigenvector, so is αx for all nonzero $\alpha \in \mathbb{C}$.

Lemma 5.2.1. The eigenvectors of $A \in \mathbb{C}^{n\times n}$ corresponding to distinct eigenvalues are linearly independent.

Proof. Suppose that $x_1, x_2, \ldots x_k$ are eigenvectors with distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Assume that these eigenvectors are not linearly independent. Then there exist scalars such that $\sum_j \alpha_j x_j = \mathbf{0}$ with not all of the $\alpha_j = 0$. If any of the α_j are zero, the corresponding terms can be dropped, yielding a linear combination of a subset of m eigenvalues, with all of the $\alpha_j \neq 0$, that yields the zero vector. Over all such linear combinations there is a subset using the smallest value of m. Denote this least value of m by p. Now assume, without loss of generality, that such a sum can be formed using the first p eigenvectors. So there exists $\alpha_j \neq 0$, $j = 1, \ldots, p$ such that

$$\sum_{j=1}^{p} \alpha_j x_j = \mathbf{0}. \tag{5.1}$$

Multiplying both sides of (5.1) by A, using the eigenvalue property, and then subtracting from this the result of multiplying both sides of (5.1) by λ_p, yields

$$\sum_{j=1}^{p-1} (\lambda_j - \lambda_p)\alpha_j x_j = \mathbf{0}.$$

Since the λ_j are distinct and $\alpha_j \neq 0$, all of the coefficients in this sum are nonzero. Hence there is a linear combination using nonzero coefficients of $p-1$ eigenvectors that is zero; a contradiction. □

It follows from Lemma 5.2.1 that if A has k distinct eigenvalues, then A has at least k linearly independent eigenvectors. However, this is only a lower bound. Depending on the particular matrix, there can be anywhere from k to n linearly independent eigenvectors.

5.2.4 Symmetric and Postive Semidefinite Matrices

Recall that a square matrix $S \in \mathbb{R}^{n\times n}$ is symmetric if $S^T = S$. The eigenvalues and eigenvectors of real symmetric matrices have some special properties.

Lemma 5.2.2. A symmetric matrix $S \in \mathbb{R}^{n\times n}$ has n real eigenvalues and n real orthonormal eigenvectors.

Proof. Let $Sx = \lambda x$ with $x \neq 0$. Then $S\bar{x} = \overline{Sx} = \bar{\lambda}\bar{x}$. Hence $\bar{x}^T Sx = \bar{\lambda}\|x\|^2$ and $\bar{x}^T Sx = \lambda\|x\|^2$. Subtracting these expressions and using $x \neq 0$, yields $\lambda = \bar{\lambda}$. Thus λ is real. It follows that x can be selected in $\mathbb{R}^n$. We prove the second claim under the simplifying assumption that S has distinct eigenvalues. Let $Sx_1 = \lambda_1 x_1$ and $Sx_2 = \lambda_2 x_2$. Then $x_2^T Sx_1 = \lambda_2 x_2^T x_1$ and $x_2^T Sx_1 = \lambda_1 x_2^T x_1$. Subtracting these expressions and using the fact that $\lambda_1 \neq \lambda_2$ yields $x_2^T x_1 = 0$. Thus $x_1 \perp x_2$. For a proof without our simplifying assumption, see Theorem 2.5.6 in Horn and Johnson. □

If we place the n orthonormal eigenvectors of S in the columns of the matrix V and place the corresponding eigenvalues on the diagonal of the diagonal matrix Λ, then $SV = V\Lambda$ and hence $S = V\Lambda V^T$.

A square matrix $P \in \mathbb{R}^{n\times n}$ with the property that for all $x \in \mathbb{R}^n$, $x^TPx \geq 0$ is said to be *positive semidefinite* (PSD). We say that P is *positive definite* (PD) if for all $x \neq 0$, $x^TPx > 0$. Without loss of generality, we will always assume P is symmetric since $x^TPx = 1/2\, x^T(P + P^T)x = x^TQx$ with $Q = 1/2(P + P^T)$ symmetric.

Here is a fundamental property of symmetric PSD and PD matrices.

Lemma 5.2.3. If $P \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite (resp. positive definite), then all the eigenvalues of P are real and nonnegative (resp. positive) and the eigenvectors of P can be selected to be real and orthonormal.

Proof. Since P is symmetric all of its eigenvalues are real and it has a set of n real ON eigenvectors. Let x be an eigenvector with eigenvalue λ. Since P is PSD, $x^TPx = x^T\lambda x = \lambda\|x\|^2 \geq 0$. Hence $\lambda \geq 0$. If P is PD and $x \neq 0$, then $x^TPx > 0$. Hence $\lambda\|x\|^2 > 0$ and thus $\lambda > 0$. □

5.2.5 Differentiable Functions $\mathbf{f}\colon \mathbb{R}^{\mathbf{n}} \to \mathbb{R}$

The Euclidean norm in $\mathbb{R}^n$ is also known as the 2-norm and is often denoted by $\|\cdot\|_2$. We will henceforth adopt this notation.

A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable if at $x \in \mathbb{R}^n$ if the exists a unique linear function $Df(x)\colon \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{h\to 0} \frac{|f(x+h) - f(x) - Df(x)h|}{\|h\|_2} = 0.$$

Here $h \in \mathbb{R}^n$, $Df(x)h \in \mathbb{R}$, and the norm in the denominator is the 2-norm in $\mathbb{R}^n$. Conceptually, $Df(x)$ is the best linear approximation to f at the point x: $f(x+h) \approx f(x) + Df(x)h$. Operationally, $Df(x)$ can be computed using partial derivatives. For this it is convenient to write $x = (x_1, x_2, \ldots, x_n)$ and $h = (h_1, h_2, \ldots, h_n)$. Then

$$Df(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

and $Df(x)$ acts on $h \in \mathbb{R}^n$ as follows:

$$Df(x)h = \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} h_j \in \mathbb{R}.$$

The *gradient* of f at x is the vector in $\mathbb{R}^n$ given by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix}^T.$$

$\nabla f(x)$ points in the direction of steepest ascent of f at the point x and $\|\nabla f(x)\|_2$ gives the rate of increase of f in this direction. We can check this by bringing in the unit norm vector $u = \nabla f(x)/\|\nabla f(x)\|_2$. Then to first order, $f(x+u) = f(x) + \nabla f(x)^T u = f(x) + \|\nabla f(x)\|_2$. So the rate of increase of f at x is $\|\nabla f(x)\|_2$ as claimed.

Example 5.2.1. Here are some simple examples.

(1) Fix $a \in \mathbb{R}^n$ and for $x \in \mathbb{R}^n$ set $f(x) = a^Tx$. This is a linear function from $\mathbb{R}^n$ to $\mathbb{R}$. In this case,

$$\begin{aligned} Df(x)(h) &= a^Th, \\ \nabla f(x) &= a. \end{aligned} \tag{5.2}$$

Notice that the derivative of a linear function is itself.

(2) Let $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ and $f(x) = x^T Ax$. Then

$$\begin{aligned} Df(x)(h) &= h^T Ax + x^T Ah = x^T(A^T + A)h, \\ \nabla f(x) &= (A^T + A)x. \end{aligned} \tag{5.3}$$

(3) Let $x \in \mathbb{R}^n$ and $f(x) = \|x\|_2^2$. Then

$$\begin{aligned} Df(x)(h) &= h^T x + x^T h = 2x^T h, \\ \nabla f(x)(h) &= 2x. \end{aligned} \tag{5.4}$$

(4) Let $x \in \mathbb{R}^n$ and $g(x) = \|x\|_2$. Then

$$\begin{aligned} Dg(x)(h) &= {}^1\!/_2(x^T x)^{-1/2}(h^T x + x^T h) = \frac{1}{\|x\|_2} x^T h, \\ \nabla f(x) &= x/\|x\|_2. \end{aligned} \tag{5.5}$$

5.2.6 Induced Norm of a Matrix

The gain of a matrix $A \in \mathbb{R}^{m\times n}$ acting on a unit norm vector $x \in \mathbb{R}^n$ is $\|Ax\|_2$. More generally, for $x \neq 0$, the gain is $\|Ax\|_2/\|x\|_2$, where in the numerator the norm is in $\mathbb{R}^m$, and in the denominator it is in $\mathbb{R}^n$. The maximum gain of A over all $x \in \mathbb{R}^n$ is then:

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \tag{5.6}$$

This is called the *induced matrix 2-norm* of A. It is induced by the Euclidean norms on $\mathbb{R}^n$ and $\mathbb{R}^m$.

To solve problem (5.6), we can restrict attention to unit norm vectors and maximize the normed squared. This yields the following equivalent problem

$$\max_{x\in\mathbb{R}^n} \quad x^T(A^T A)x \tag{5.7}$$

$$\text{s.t.} \quad x^T x = 1. \tag{5.8}$$

A vector $x \in \mathbb{R}^n$ that solves this problem also provides a solution to (5.6).

Theorem 5.2.2 (Horn and Johnson, 4.2.2)**.** Let $P \in \mathbb{R}^{n\times n}$ be a symmetric positive semidefinite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The problem

$$\begin{aligned} \max_{u\in\mathbb{R}^n} \quad & u^T Pu \\ \text{s.t.} \quad & u^T u = 1 \end{aligned} \tag{5.9}$$

has the optimal value λ_1 and this is achieved if and only if u is a unit norm eigenvector of P for λ_1. If $\lambda_1 > \lambda_2$, the solution is unique up the sign of u.

Proof. We want to maximize $x^T Px$ subject to $x^T x = 1$. Bring in a Lagrange multiplier μ and form the Lagrangian $L(x, \mu) = x^T Px + \mu(1 - x^T x)$. Taking the derivative of this expression with respect to x and setting this equal to zero yields the necessary condition $Px = \mu x$. Hence μ must be a eigenvalue of P with x a corresponding eigenvector normalized so that $x^T x = 1$. For such x, $x^T Px = \mu x^T x = \mu$. Hence the maximum achievable value of the objective is λ_1 and this is achieved when u a corresponding eigenvector of P. Conversely, if u is any unit norm eigenvector of P for λ_1, then $u^T Pu = \lambda_1$ and hence u is a solution. □

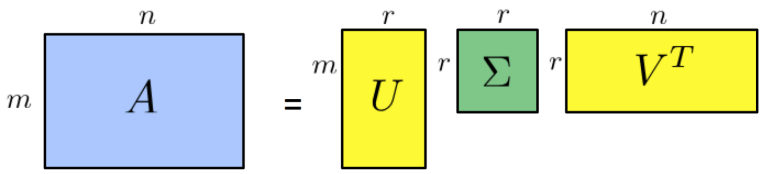

Figure 5.1: A visualization of the matrices in the compact SVD.

Using Theorem 5.2.2, a solution of (5.6) is obtained by selecting x to be a unit norm eigenvector of A^TA for its largest eigenvalue. Hence

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)} \tag{5.10}$$

Because of this connection with eigenvalues, the induced matrix 2-norm is sometimes also called the *spectral norm*.

It is easy to check that the induced norm is a norm (exercise). Moreover, it has the following additional properties.

Lemma 5.2.4. Let A, B be matrices of appropriate size and $x \in \mathbb{R}^n$. Then

1) $\|Ax\|_2 \leq \|A\|_2\|x\|_2$;

2) $\|AB\|_2 \leq \|A\|_2 \, \|B\|_2$.

Proof. Exercise. □

Important: The induced matrix 2-norm and the matrix Euclidean norm are distinct norms on $\mathbb{R}^{m\times n}$. Recall, the Euclidean norm on $\mathbb{R}^{m\times n}$ is called the Frobenius norm and is denoted by $\|A\|_F$.

5.3 Singular Value Decomposition

We first present the main SVD result in what is called the compact form. We then give interpretations of the SVD and indicate an alternative version known as the full SVD. After these discussions, we turn our attention to the ideas and constructions that form the foundation of the SVD.

Theorem 5.3.1 (Singular Value Decomposition)**.** Let $A \in \mathbb{R}^{m\times n}$ have rank $r \leq \min\{m, n\}$. Then there exist $U \in \mathbb{R}^{m\times r}$ with $U^TU = I_r$, $V \in \mathbb{R}^{n\times r}$ with $V^TV = I_r$, and a diagonal matrix $\Sigma \in \mathbb{R}^{r\times r}$ with diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$, such that

$$A = U\Sigma V^T = \sum_{j=1}^{r} \sigma_j u_j v_j^T.$$

The positive scalars σ_j are called the *singular values* of A. The r orthonormal columns of U are called the *left* or *output singular vectors* of A, and the r orthonormal columns of V are called the *right* or *input singular vectors* of A. The conditions $U^TU = I_r$ and $V^TV = I_r$ indicated that U and V have orthonormal columns. In general, U and V are not orthogonal matrices, i.e., $UU^T \neq I_m$ and $VV^T \neq I_n$. Notice also that the theorem does not claim that U and V are unique. We discuss this issue later in the chapter. The compact SVD decomposition is illustrated in Fig. 5.1.

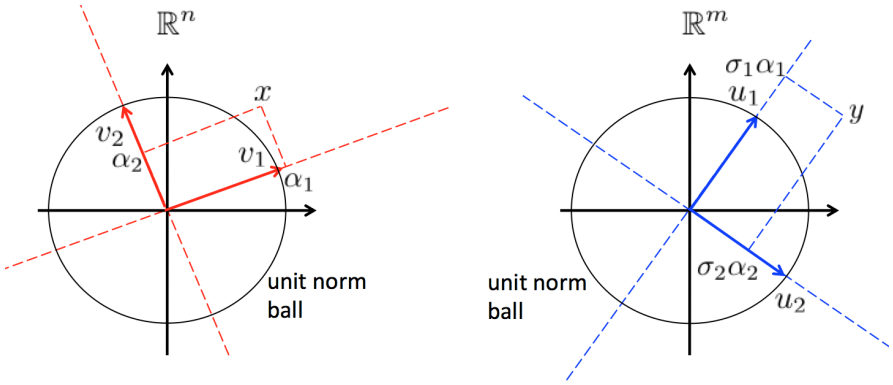

Figure 5.2: A visualization of the three operational steps in the compact SVD. The projection of $x \in \mathbb{R}^n$ onto $\mathcal{N}(A)^\perp$ is represented in terms of the basis v_1, v_2. Here $x = \alpha_1 v_1 + \alpha_2 v_2$. These coordinates are scaled by the singular values. Then the scaled coordinates are transferred to the output space $\mathbb{R}^m$ and used to the form the result $y = Ax$ as the linear combination $y = \sigma_1\alpha_1 u_1 + \sigma_2\alpha_2 u_2$.

Corollary 5.3.1. The matrices U and V in the compact SVD have the following additional properties:

a) The columns of U form an orthonormal basis for the range of A.

b) The columns of V form an orthonormal basis for $\mathcal{N}(A)^\perp$.

Proof. a) Writing $Ax = U(\Sigma V^T x)$ shows that $Ax \in \mathcal{R}(U)$ and hence that $\mathcal{R}(A) \subseteq \mathcal{R}(U)$. Let $u \in \mathcal{R}(U)$. Then for some $z \in \mathbb{R}^r$, $u = Uz = U\Sigma V^T V(\Sigma^{-1}z) = A(V^T\Sigma^{-1}z)$. Hence $\mathcal{R}(U) \subseteq \mathcal{R}(A)$.
b) By taking transposes and using part a), the columns of V form an ON basis for the range of A^T. Using $\mathcal{N}(A)^\perp = \mathcal{R}(A^T)$ yields the desired result. □

The above observations lead to the following operational interpretation of the SVD. For $x \in \mathbb{R}^n$, the operation $V^T x$ gives the coordinates with respect to V of the orthogonal projection of x onto the subspace $\mathcal{N}(A)^\perp$ (the orthogonal projection is $\hat{x} = VV^T x$). These r coordinates are then individually scaled using the r diagonal entries of Σ. Finally, we synthesize the output vector by using the scaled coordinates and the ON basis U for $\mathcal{R}(A)$: $y = U(\Sigma V^T x)$. So the SVD has three steps: (1) An analysis step: $V^T x$, (2) A scaling step: $\Lambda(V^T x)$, and (3) a synthesis step: $U(\Sigma V^T x)$. In particular, when $x = v_k$, $y = Ax = \sigma_k u_k$, $k = 1, \ldots, r$. So the r ON basis vectors for $\mathcal{N}(A)^\perp$ are mapped to scaled versions of corresponding ON basis vectors for $\mathcal{R}(A)$. This is illustrated in Fig. 5.2. Notice that restricted to $\mathcal{N}(A)^\perp$, the map $A\colon \mathcal{N}(A)^\perp \to \mathcal{R}(A)$ is one-to-one and onto and hence invertible.

5.3.1 Singular Values and Norms

The Frobenius norm and the induced matrix 2-norm of a matrix $A \in \mathbb{R}^{m\times n}$ are both related to the singular values of A. First note that

$$\|Ax\|^2 = \|U\Sigma V^T x\|^2 = x^T(V\Sigma^2 V^T)x.$$

To maximize this expression we select x to be a unit norm eigenvector of $(V\Sigma^2 V^T)$ with maximum eigenvalue. Hence we use $x = v_1$ and achieve $\|Ax\|^2 = \sigma_1^2$. So the input direction with the most gain is v_1, this appears in the output in the direction u_1, and the gain is σ_1: $Av_1 = \sigma_1 u_1$. This is vizuallized in Fig. 5.3. We conclude that the induced 2-norm of A equals the maximum singular value of A:

$$\|A\|_2 = \sigma_1. \tag{5.11}$$

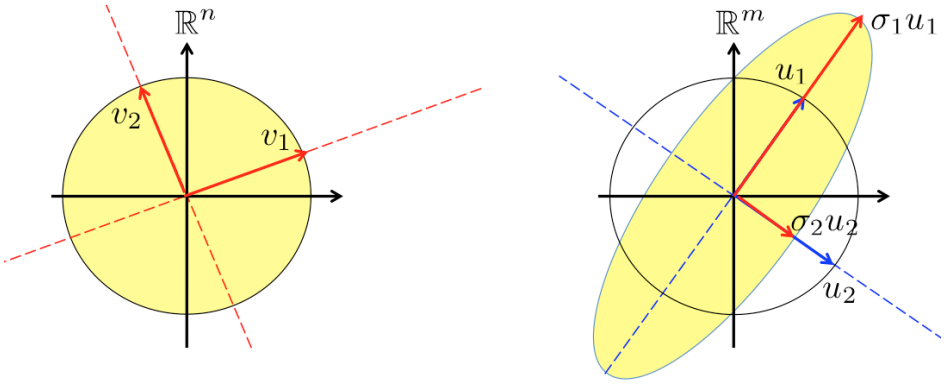

Figure 5.3: A visualization of the action of A on the unit sphere in $\mathbb{R}^n$ in terms of its SVD.

Now we express the Frobenius norm of matrix in terms of its singular values. To do so, let $A \in \mathbb{R}^{m\times n}$ have rank r and compact SVD $A = U\Sigma V^T = \sum_{j=1}^r \sigma_j u_j v_j^T$. In this form we see that the SVD expresses A as a positive linear combination of the rank one matrices $u_j v_j^T$, $j = 1, \ldots, r$. These matrices form an orthonormal set in $\mathbb{R}^{m\times n}$:

$$< u_k v_k^T, u_j v_j^T > = \text{trace}(v_k u_k^T u_j v_j^T) = \text{trace}(u_k^T u_j v_j^T v_k) = \begin{cases} 0, & \text{if } j \neq k; \\ 1, & \text{if } j = k. \end{cases}$$

So the SVD is selecting an orthonormal basis of rank one matrices $\{u_j v_j^T\}_{j=1}^r$ specifically adapted to A, and expressing A as a positive linear combination of this basis. With these insights, we can apply Pythagorous' Theorem to the expression $\|A\|_F^2 = \| \sum_{j=1}^r \sigma_j u_j v_j^T \|_F^2$ to obtain:

$$\|A\|_F = \left(\sum_{j=1}^r \sigma_j^2 \right)^{1/2}. \tag{5.12}$$

So the Frobenius norm of A is the Euclidean norm of the vector of its singular values.

The following Theorem will be useful in the discussion of PCA in the next chapter.

Theorem 5.3.2. Let $A, B \in \mathbb{R}^{m\times n}$, $Q \in \mathcal{O}_m$, and $R \in \mathcal{O}_n$. Let A and B have compact SVDs $A = U_A \Sigma_A V_A^T$ and $U_B \Sigma_B V_B^T$. Then

$$\|A - QBR\|_F \geq \|\Sigma_A - \Sigma_B\|_F, \tag{5.13}$$

with equality when Q and R are selected to make $QU_B = U_A$ and $R^T B_V = V_A$.

Proof. Horn and Johnson, p436. □

5.3.2 The Full SVD

There is a second version of the SVD that is often convenient in various proofs involving the SVD. Often this second version is just called the SVD. However, to emphasize its distinctness from the equally useful compact SVD, we refer to it as a *full SVD*.

The basic idea is very simple. Let $A = U_c \Sigma_c V_c^T$ be a compact SVD with $U_c \in \mathbb{R}^{m\times r}$, $V_c \in \mathbb{R}^{n\times r}$, and $\Sigma_c \in \mathbb{R}^{r\times r}$. To U_c we add an orthonormal basis for $\mathcal{R}(U_c)^\perp$ to form the orthogonal matrix $U = \begin{bmatrix} U_c & U_c^\perp \end{bmatrix} \in \mathbb{R}^{m\times m}$. Similarly, to V_c we add an orthonormal basis for $\mathcal{R}(V_c)^\perp$ to form the orthogonal

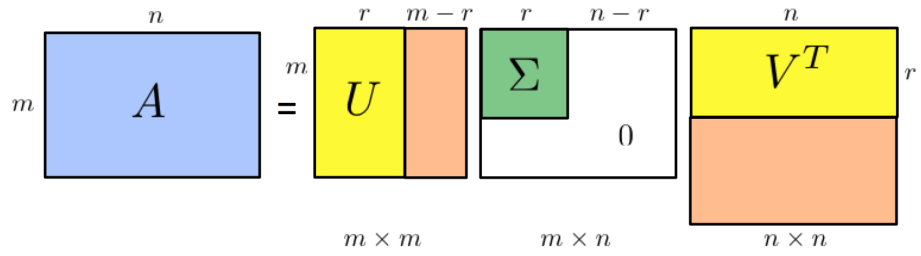

Figure 5.4: A visualization of the matrices in the full SVD.

matrix $V = \begin{bmatrix} V_c & V_c^\perp \end{bmatrix} \in \mathbb{R}^{n\times n}$. To ensure that these extra columns in U and V do not interfere with the factorization of A, we form $\Sigma \in \mathbb{R}^{m\times n}$ by padding Σ_c with zero entries:

$$\Sigma = \begin{bmatrix} \Sigma_c & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{bmatrix}.$$

We then have a full SVD factorization $A = U\Sigma V^T$. The utility of the full SVD derives from U and V being orthogonal (hence invertible) matrices. The full SVD is illustrated in Fig. 5.4.

If P is a symmetric positive semidefinite matrix, a full SVD of P is simply an eigendecomposition of P: $U\Sigma V^T = Q\Sigma Q^T$, where Q is the orthogonal matrix of eigenvectors of P. In this sense, the SVD extends the eigendecomposition by using different orthonormal sets of vectors in the input and output spaces.

5.4 Inner Workings of the SVD

We now give a quick overview of the origins of the matrices U, V and Σ in a SVD. Let $A \in \mathbb{R}^{m\times n}$ have rank r. So the range of A has dimension r and the nullspace of A has dimension $n - r$.

Let $B = A^T A \in \mathbb{R}^{n\times n}$. Since B is a symmetric positive semi-definite (PSD) matrix, it has nonnegative eigenvalues and full set of orthonormal eigenvectors. Order the eigenvalues in decreasing order: $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2 \geq 0$ and let v_j denote the eigenvector for σ_j^2. So

$$Bv_j = \sigma_j^2 v_j, \quad j = 1, \ldots, n.$$

Noting that $Ax = 0$ if and only if $Bx = 0$ (Exercise 5.3) we see that the null space of B also has dimension $n-r$. It follows that $n-r$ of the eigenvectors of B must lie in $\mathcal{N}(A)$ and r must lie in $\mathcal{N}(A)^\perp$. Hence

$$\sigma_1^2 \geq \cdots \geq \sigma_r^2 > 0 \quad \text{and} \quad \sigma_{r+1}^2 = \cdots = \sigma_n^2 = 0,$$

with $v_1, \ldots, v_r$ an orthonormal basis for $\mathcal{N}(A)^\perp$.

Now consider $C = AA^T \in \mathbb{R}^{m\times m}$. This is also symmetric and PSD. Hence C has nonnegative eigenvalues and a full set of orthonormal eigenvectors. Order the eigenvalues in decreasing order: $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_m \geq 0$ and let u_j denote the eigenvector for λ_j^2. So

$$Cu_j = \lambda_j^2 u_j, \quad j = 1, \ldots, m.$$

Since $\mathcal{R}(A^T) = \mathcal{N}(A)^\perp$, the dimension of $\mathcal{R}(A^T)$ is r, and that of $\mathcal{N}(A^T)$ is $m - r$. By the same reasoning as above, $m - r$ of the eigenvectors of C must lie in $\mathcal{N}(A^T)$ and r must lie in $\mathcal{R}(A)$. Hence

$$\lambda_1^2 \geq \cdots \geq \lambda_r^2 > 0 \quad \text{and} \quad \lambda_{r+1}^2 = \cdots = \lambda_m^2 = 0,$$

with $u_1, \ldots, u_r$ an orthonormal basis for $\mathcal{R}(A)$.

Now we show a relationship between σ_j^2, λ_j^2 and the corresponding eigenvectors v_j, u_j, for $j = 1, \ldots, r$. First consider $Bv_j = \sigma_j^2 v_j$ with $\sigma_j^2 > 0$. We have

$$C(Av_j) = (AA^T)(Av_j) = A(A^T A v_j) = A(Bv_j) = \sigma_j^2 (Av_j).$$

So either $Av_j = 0$, or Av_j is an eigenvector of C with eigenvalue σ_j^2. The first case, $Av_j = 0$ contradicts $A^T A v_j = \sigma_j^2 v_j$ with $\sigma_j^2 > 0$ since $Av_j = 0$ implies $(A^T A)v_j = 0$. Hence Av_j must be an eigenvector of C with eigenvalue σ_j^2. Assume for simplicity, that the positive eigenvalues of $A^T A$ and AA^T are distinct. Then for some k, with $1 \leq k \leq r$:

$$\sigma_j^2 = \lambda_k^2 \quad \text{and} \quad Av_j = \alpha u_k, \text{ with } \alpha > 0.$$

We can take $\alpha > 0$ by swapping $-u_k$ for u_k if necessary. Using this result we find

$$v_j^T B v_j = \begin{cases} \sigma_j^2 v_j^T v_j = \sigma_j^2; \\ (Av_j)^T (Av_j) = \alpha^2 u_j^T u_j = \alpha^2. \end{cases}$$

So we must have $\alpha = \sigma_j$ and

$$Av_j = \sigma_j u_k.$$

Doing the same analysis for $Cu_k = \lambda_k^2 u_k$ with $\lambda_k^2 > 0$, yields

$$B(A^T u_k) = (A^T A)(A^T u_k) = A^T (AA^T u_k) = \lambda_k^2 (A^T u_k).$$

Since $\lambda_k^2 > 0$, we can't have $A^T u_k = 0$. So $A^T u_k$ is an eigenvector of $A^T A$ with eigenvalue λ_k^2. Under the assumption of distinct nonzero eigenvalues, this implies that for some p with $1 \leq p \leq r$,

$$\lambda_k^2 = \sigma_p^2 \quad \text{and} \quad A^T u_k = \beta v_p, \text{ some } \beta \neq 0.$$

Using this expression to evaluate $u_k^T C u_k$ we find $\lambda_k^2 = \beta^2$. Hence $\beta^2 = \lambda_k^2 = \sigma_p^2$ and $A^T u_k = \beta v_p$.

We now have two ways to evaluate $A^T A v_j$:

$$A^T A v_j = \begin{cases} \sigma_j^2 v_j & \text{by definition;} \\ \alpha A^T u_k = \alpha\beta v_p. & \text{using the above analysis.} \end{cases}$$

Equating these answers gives $j = p$ and $\alpha\beta = \sigma_j^2$. Since $\alpha > 0$, it follows that $\beta > 0$ and $\alpha = \sigma_j = \lambda_j = \beta$. Thus $Av_j = \sigma_j u_j$, $j = 1, \ldots, r$. Written in matrix form this is almost the compact SVD:

$$A \begin{bmatrix} v_1 & \ldots & v_r \end{bmatrix} = \begin{bmatrix} u_1 & \ldots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}.$$

From this we deduce that $AVV^T = U\Sigma V^T$. VV^T computes the orthogonal projection of x onto $\mathcal{N}(A)^\perp$. Hence for every $x \in \mathbb{R}^n$, $AVV^T x = Ax$. Thus $AVV^T = A$, and we have $A = U\Sigma V^T$.

Finally note that $\sigma_j = \sqrt{\lambda_j(A^T A)} = \sqrt{\lambda_j(AA^T)}$, $j = 1, \ldots, r$. So the singular values are always unique. If the singular values are distinct, the SVD is unique up to sign interchanges between the u_j and v_j. But this still leaves the representation $A = \sum_{j=1}^r \sigma_j u_j v_j^T$ unique. If the singular values are not distinct, then U and V are not unique. For example, $I_n = U I_n U^T$ for every orthogonal matrix U.

5.5 Notes

For more detailed reading about the SVD see Horn and Johnson [1].

5.6 Exercises

Preliminaries:

Exercise 5.1. Let $A \in \mathbb{R}^{m\times n}$. The rank of A is the dimension of $\mathcal{R}(A)$. Show that: (a) the rank of A equals the number of linearly independent columns of A; and (b) the rank of A equals the number of linearly independent rows of A.

Exercise 5.2. Let $A \in \mathbb{R}^{m\times n}$ have rank r. Show that

(a) $\dim \mathcal{N}(A) = n - r$

(b) $\dim \mathcal{R}(A)^{\perp} = m - r$

(c) $\dim \mathcal{N}(A)^{\perp} = \dim \mathcal{R}(A)$

(d) $\dim \mathcal{N}(A) + \dim \mathcal{R}(A) = n$

Exercise 5.3. Let $A \in \mathbb{R}^{m\times n}$ and $B = A^T A$. Show that $\mathcal{N}(A) = \mathcal{N}(B)$.

Induced 2-norm:

Exercise 5.4. Show that the induced 2-norm satisfies the properties of a norm.

Exercise 5.5. Prove Lemma 5.2.4.

Basic SVD Properties:

Exercise 5.6. Let P be a real symmetric $n \times n$ matrix. Show how to form an SVD factorization of P from its eigendecomposition $P = Q\Omega Q^T$. Here the columns of Q are eigenvectors of P and Ω is a diagonal matrix with the corresponding eigenvalues on the diagonal.

Exercise 5.7. Show that the eigenvalues of a symmetric PSD matrix are also its singular values.

Exercise 5.8. For $\alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{m\times n}$, show that $\sigma_i(\alpha A) = |\alpha|\sigma_i(A)$, $i = 1, \ldots, q = \min(m, n)$.

Exercise 5.9. For $A, B \in \mathbb{R}^{m\times n}$. Show that $\sigma_1(A + B) \leq \sigma_1(A) + \sigma_1(B)$.

Exercise 5.10. For $A \in \mathbb{R}^{m\times n}$ and $Q \in \mathcal{O}_m$, $R \in \mathcal{O}_n$, show that the singular values of QAR are the same as those of A. Hence singular values are invariant under orthogonal transformations.

Exercise 5.11. For any $A \in \mathbb{R}^{m\times n}$, $Q \in \mathcal{O}_m$, $R \in \mathcal{O}_n$ show that $\|QAR\|_F = \|A\|_F$. Thus the Frobenius norm is invariant under orthogonal transformations.

Exercise 5.12. Show that the induced 2-norm of $A \in \mathbb{R}^{m\times n}$ is invariant under orthogonal transformations.

Exercise 5.13. Let $A \in \mathbb{R}^{m\times n}$ have rank r and compact SVD $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$. For $k = 1, \ldots, r$, let $\mathcal{U}_k = \text{span}\{u_i v_i^T\}_{i=1}^{k}$, and $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$. Clearly $A_k \in \mathcal{U}_k$ and $A_r = A$. Show the following:

(a) A_k has rank k.

(b) With respect to the Frobenius norm, A_k is the closest point in $\mathcal{U}_k$ to A.

(c) With respect to the Frobenius norm, A_k is the closest rank k matrix to A.

Inverse Maps and the SVD:

Exercise 5.14. Let $A \in \mathbb{R}^{n\times n}$ be a square invertible matrix with SVD $A = U\Sigma V^T$. Show that $A^{-1} = V\Sigma^{-1}U^T$.

Exercise 5.15. The Moore-Penrose pseudo-inverse of a matrix $A \in \mathbb{R}^{m\times n}$ is the unique matrix $A^+ \in \mathbb{R}^{n\times m}$ satisfying the following four properties:
a) $A(A^+A) = A$,
b) $(A^+A)A^+ = A^+$,
c) $(A^+A)^T = A^+A$,
d) $(AA^+)^T = AA^+$.
Let A have compact SVD $A = U\Sigma V^T$. Show that $A^+ = V\Sigma^{-1}U^T$. Give an interpretation of A^+ in terms of $\mathcal{N}(A)$, $\mathcal{N}(A)^\perp$ and $\mathcal{R}(A)$.

Exercise 5.16. Let $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$ be given. We want to find a solution x of the linear equations $A^TAx = A^Ty$. Show that if $A = U\Sigma V^T$ is a compact SVD of A, then a solution is $x^\star = V\Sigma^{-1}U^Ty$ and $x^\star \in \mathcal{N}(A)^\perp$.

Maximizing Inner Products:

Exercise 5.17. Let $\Sigma, \Lambda \in \mathbb{R}^{n\times n}$ be diagonal matrices with nonnegative diagonal entries. We want to find permutation matrices $P, Q \in \mathbb{R}^{n\times n}$ to maximize $<\Sigma, P\Lambda Q>$.

(a) Show that optimal permutation matrices $P^\star$, $Q^\star$ must exist.

(b) Show that $P^\star\Lambda Q^\star$ must be diagonal.

(c) In light of (b), show that the problem can be simplified as follows. For $x, y \in \mathbb{R}^n$ with nonnegative entries, find a permutation matrix $P \in \mathbb{R}^{n\times n}$ that maximizes the inner product $<x, Py> = x^TPy$.

(d) Solve the problem in part (c). You can initially assume that $x(1) \geq x(2) \geq \cdots \geq x(n)$. Once you have a solution in this case, show how to remove the above assumption.

Exercise 5.18. Let Σ be an $n \times n$ diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. What orthogonal matrices $W \in \mathcal{O}_n$ maximize the inner product $<W, \Sigma>$?

Exercise 5.19. For $A \in \mathbb{R}^{n\times n}$, find $Q \in \mathcal{O}_n$ to maximize the inner product $<Q, A>$ and determine the maximum value. In addition, show that $Q \in \mathcal{O}_n$ maximizes $<Q, A>$ iff Q^TA is symmetric PSD.

Exercise 5.20. For given $A, B \in \mathbb{R}^{m\times n}$, find an orthogonal matrix $Q \in \mathcal{O}_m$ to maximize the inner product $<A, QB>$. This Q "rotates" the columns of B to maximize the inner product with A.

Exercise 5.21. For given $A, B \in \mathbb{R}^{m\times n}$, we seek $Q \in \mathcal{O}_m$ to minimize $\|A - QB\|_F^2$. So we want to find an orthogonal matrix $Q \in \mathcal{O}_m$ that "rotates" the columns of B to best match A as measured by the sum of squared errors. Show that this is equivalent to finding $Q \in \mathcal{O}_m$ to maximize the inner product $<A, QB>$, and determine the minimum achievable value.

Exercise 5.22. For given $A \in \mathbb{R}^{m\times n}$ we want to find $B \in \mathbb{R}^{m\times n}$ that maximizes $<A, B>$ subject to $\|B\|_2 \leq 1$. Here B is constrained by placing an upper bound on its gain (maximum singular value). Let and $q = \min(m, n)$. Show that the maximum value is $\sum_{i=1}^{q} \sigma_i(A)$ and find a solution B.

Exercise 5.23. For $A, B \in \mathbb{R}^{m\times n}$, we seek orthogonal matrices $Q \in \mathcal{O}_m$ and $R \in \mathcal{O}_n$ to minimize $\|A - QBR\|_F^2$. Let $A = U_A\Sigma_A V_A^T$ and $B = U_B\Sigma_B V_B^T$ be full singular value decompositions. For $Q \in \mathcal{O}_m$ and $R \in \mathcal{O}_n$, show the following claims:

(a) $\min_{Q,R} \|A - QBR\|_F^2 = \mathrm{trace}(\Sigma_A^2) + \mathrm{trace}(\Sigma_B^2) - 2\max_{Q,R} \mathrm{trace}(\Sigma_A^T Q \Sigma_B R)$.

(b) $\min_{Q,R} \|A - QBR\|_F^2 = \|\Sigma_A - \Sigma_B\|_F^2$.

Exercise 5.24. Let $A, B \in \mathbb{R}^{m\times n}$ and $A = U_A\Sigma_A V_A^T$, $B = U_B\Sigma_B V_B^T$ be full SVDs. Show that $\|A - B\|_F \geq \|\Sigma_A - \Sigma_B\|_F$.

The Nuclear Norm:

Exercise 5.25. For $A \in \mathbb{R}^{m\times n}$ and $q = \min(m, n)$, define the nuclear norm by $\|A\|_* = \sum_{i=1}^{q} \sigma_i(A)$. Show that $\|\cdot\|_*$ is a norm on $\mathbb{R}^{m\times n}$.

Exercise 5.26. Show that the nuclear norm is invariant under orthogonal transformations.

Exercise 5.27. Show that $\|A\|_* = \mathrm{trace}(\sqrt{A^T A})$.

Miscellaneous:

Exercise 5.28 (Idempotent But Not Symmetric)**.** Let $P \in \mathbb{R}^{n\times n}$ have rank r and compact SVD $P = U\Sigma V^T$. If $P^2 = P$, show that either $r = n$ and $P = I_n$ or $r < n$ and $P = VV^T + V_0V^T$ where the columns of V_0 lie in $\mathcal{R}(V)^\perp$. Choosing $V_0 = \mathbf{0}$, yields a projection matrix $P = VV^T$. But choosing $V_0 \neq \mathbf{0}$, yields an idempotent matrix P that is not symmetric.

Exercise 5.29. Let $A \in \mathbb{R}^{n\times n}$ have a compact SVD $U_A\Sigma_A V_A^T$. Show that $\mathrm{trace}(A) \leq \mathrm{trace}(\Sigma_A)$.

Chapter 6

Principal Component Analysis

Given a set of data $\{x_j \in \mathbb{R}^n\}_{j=1}^p$, and an integer $q < n$, we ask if there is a q-dimensional subspace onto which we can project the data so that the sum of the squared norms of the residuals is minimized. It turns out that for every $1 \leq q < n$, there is indeed a subspace that minimizes this metric. If the resulting approximation is reasonably accurate, then the original data lies approximately on a q-dimensional subspace in $\mathbb{R}^n$. Hence, at the cost of small approximation error, we gain the benefit of reducing the dimension of the data to q. This is an example of *linear dimensionality reduction*.

This idea is connected with the question of how is the data spread out about its sample mean. Directions in which the data does not have significant variation could be eliminated allowing the data to be represented in a lower dimensional subspace. This leads to a core method of dimensionality reduction known as Principal Component Analysis (PCA). It selects a projection of the data onto q-dimensional subspace that maximizes the captured variance of the original data. This subspace is the same subspace described above that minimizes the sum of squared norms of the resulting residuals.

To make this more concrete assume that the data has been centered and let x_j denote that the j-th (centered) data point. Select a subspace $\mathcal{U}$ of dimension q and let the columns of $U \in \mathcal{O}_{n\times q}$ be an orthonormal basis for $\mathcal{U}$. Then the projection of $x_j \in \mathbb{R}^n$ onto $\mathcal{U}$ is given by $\hat{x}_j = UU^T x_j \in \mathbb{R}^n$. Notice that this does not reduce the dimension of the data; it simply maps data points onto a q-dimensional subspace in $\mathbb{R}^n$. There is an error (the residual) associated with this projection $x_j = \hat{x}_j + r_j$. Our intent is to replace x_j with $\hat{x}_j$ and suffer the loss of any information in r_j. So the projection is a lossy operation. Nevertheless, there may still be sufficient information in the projected data points to accomplish the task at hand. The advantage is that we can work a reduced dimension space. This is achieved by noting that instead of working with $\hat{x}_j$ we can equivalently work its coordinates with $c_j \in \mathbb{R}^q$ with respect to the basis U. So $\hat{x}_j = Uc_j$. Hence $c_j = U^T\hat{x}_j = U^T x_j$. There is no loss of information in working with the c_j in place of the $\hat{x}_j$ and its the advantage of reducing the dimension of the point points from n to q. This approach is called linear dimensionality reduction. Different methods of linear dimensionality reduction arise by selecting the subspace $\mathcal{U}$ in different ways.

6.1 Preliminaries

6.1.1 Centering the Data

Recall that the *sample mean* of a set of data $\{x_j \in \mathbb{R}^n\}_{j=1}^p$ is the vector $\mu = {}^1\!/_p \sum_{j=1}^p x_j$ and that we center the data by subtracting μ from each x_j, forming $z_j = x_j - \mu$. The centered data has zero sample

mean:

$$1/p \sum_{j=1}^{p} z_j = 1/p \sum_{j=1}^{p} (x_j - \mu) = \mu - \mu = \mathbf{0}.$$

The centering operation can be expressed in matrix form as follows. Form the data into the matrix $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$. Then $\mu = 1/p X\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^p$ denotes the vector of all 1's. Let Z denote the corresponding matrix of centered data and $u = (1/\sqrt{p})\mathbf{1}$. Then

$$Z = X - \mu \mathbf{1}_p^T = X - 1/p X \mathbf{1}_p \mathbf{1}_p^T = X(I_p - 1/p \mathbf{1}_p \mathbf{1}_p^T) = X(I_p - uu^T).$$

From this point forward we assume that the data has been centered.

6.1.2 Parameterizing the Family of q-Dimensional Subspaces

We want to learn the "best" subspace $\mathcal{U}$ of dimension $q \leq n$ that adequately represents the data. To do so it is convenient to have a parameterization of q-dimensional subspaces of $\mathbb{R}^n$. A subspace $\mathcal{U} \subseteq \mathbb{R}^n$ of dimension $q \leq n$ can be represented by an orthonormal basis for $\mathcal{U}$. However, this representation is not unique since there are infinitely many orthonormal bases for $\mathcal{U}$. Any such basis contains q orthonormal vectors and these can be arranged into the columns of a $n \times q$ matrix $U = [u_1, \ldots, u_q] \in \mathbb{R}^{n \times q}$ with $U^T U = I_q$.

Let $U_1, U_2 \in \mathbb{R}^{n \times q}$ be two orthonormal bases for the same q-dimensional subspace $\mathcal{U}$. Since U_1 is a basis for $\mathcal{U}$ and every column of U_2 lies in $\mathcal{U}$, there must exist a matrix $Q \in \mathbb{R}^{q \times q}$ such that $U_2 = U_1 Q$. It follows that $Q = U_1^T U_2$. Using $U_1 U_1^T U_2 = U_2$ and $U_2 U_2^T U_1 = U_1$, we then have

$$\begin{aligned} Q^T Q &= U_2^T U_1 U_1^T U_2 = U_2^T U_2 = I_k, \text{ and} \\ QQ^T &= U_1^T U_2 U_2^T U_1 = U_1^T U_1 = I_k. \end{aligned}$$

Hence $Q \in \mathcal{O}_q$. So any two orthonormal basis representations U_1, U_2 of $\mathcal{U}$ are related by a $q \times q$ orthogonal matrix Q: $U_2 = U_1 Q$ and $U_1 = U_2 Q^T$.

6.2 An Optimal Projection Subspace

Our goal here is to find a q-dimensional subspace $\mathcal{U}$ such that the orthogonal projection of the data onto $\mathcal{U}$ minimizes the sum of squared norms of the residuals. Assuming such a subspace exists, we call it an *optimal projection subspace* of dimension q. If the columns of $U \in \mathbb{R}^{n \times q}$ form an orthonormal basis for $\mathcal{U}$, then the matrix of projected data is $\hat{X} = UU^T X$ and the corresponding matrix of residuals is $(I - UU^T)X$. Hence we seek to solve:

$$\begin{aligned} \min_{U \in \mathbb{R}^{n \times q}} \quad & \|X - UU^T X\|_F^2 \\ \text{s.t.} \quad & U^T U = I_q. \end{aligned} \tag{6.1}$$

The solution of this problem can't be unique since if U is a solution so is UQ for every $Q \in \mathcal{O}_q$. These solutions correspond to different parameterizations of the same subspace. It is also of interest to determine if two distinct subspaces could both be optimal projection subspaces of dimension q.

Using standard equalities, the objective function of (6.1) can be rewritten as

$$\|X - UU^T X\|_F^2 = \text{trace}(X^T(I - UU^T)(I - UU^T)X) = \text{trace}(XX^T) - \text{trace}(U^T XX^T U).$$

Hence, letting $P \in \mathbb{R}^{n\times n}$ denote the symmetric positive semidefinite matrix XX^T, we can equivalently solve the following problem:

$$\begin{aligned} \max_{U\in\mathbb{R}^{n\times q}} \quad & \text{trace}(U^T P U) \\ \text{s.t.} \quad & U^T U = I_q. \end{aligned} \tag{6.2}$$

Problem (6.2) is well known. The simplest version with $q = 1$ was covered in Theorem 5.2.2. Below we give a generalization of that theorem. The proof uses a result we have discussed but not proved.

Theorem 6.2.1 (Horn and Johnson, 4.3.18)**.** Let the symmetric positive semidefinite matrix $P \in \mathbb{R}^{n\times n}$ have eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then the optimal value of problem (6.2) is $\sum_{j=1}^{q} \lambda_j$. Moreover, this is achieved if the columns of U are q orthonormal eigenvectors for the largest q eigenvalues of P.

Proof. Let $P = V\Sigma V^T$ be an eigen-decomposition of P with $V \in \mathcal{O}_n$ and $\Sigma \in \mathbb{R}^{n\times n}$ a diagonal matrix with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ listed in decreasing order down the diagonal. We want to maximize $\text{trace}(U^T V\Sigma V^T U) = \text{trace}(\Sigma V^T U U^T V) = \text{trace}(\Sigma W W^T)$, where $W = V^T U \in \mathcal{O}_{n\times q}$. This is equivalent to maximizing $<\Sigma, WW^T>$ by choice of $W \in \mathcal{O}_{n\times q}$, then setting $U = VW$.

The maximization of $<\Sigma, WW^T>$ can be solved as follows. Let $Z \in \mathcal{O}_{n,n-q}$ be an orthonormal basis for $\mathcal{R}(W)^\perp$. Then the matrices Σ and WW^T have the following full singular value decompositions

$$\Sigma = \begin{bmatrix} I_q & \mathbf{0} \\ \mathbf{0} & I_{n-q} \end{bmatrix} \Sigma \begin{bmatrix} I_q & \mathbf{0} \\ \mathbf{0} & I_{n-q} \end{bmatrix}^T \quad \text{and} \quad WW^T = \begin{bmatrix} W & Z \end{bmatrix} \begin{bmatrix} I_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-q} \end{bmatrix} \begin{bmatrix} W^T \\ Z^T \end{bmatrix}.$$

It is a standard result that if we are free to select the left and right singular vectors of a matrix B, then the inner product $<A, B>$ is maximized when the left and right singular vectors of B are chosen to equal the left and right singular vectors of A, respectively. Hence selecting $W = \begin{bmatrix} I_q & \mathbf{0} \end{bmatrix}^T$ maximizes the inner product $<\Sigma, WW^T>$. This gives $U = V_q$, where V_q is the matrix of the first q columns of V and results in the optimal objective value $\sum_{j=1}^{q} \lambda_j$. □

It follows from Theorem 6.2.1, that a solution $U^\star$ to (6.2) is obtained by selecting the columns of $U^\star$ to be a set orthonormal eigenvectors of $P = XX^T$ corresponding to its q largest eigenvalues. Working backwards, we see that $U^\star$ is then also a solution to (6.1). In both cases, there is nothing special about $U^\star$ beyond the fact that it spans $\mathcal{U}^\star$. Any basis of the form $U^\star Q$ with $Q \in \mathcal{O}_k$ spans the same optimal subspace $\mathcal{U}^\star$. We also note that the subspace $\mathcal{U}^\star$ may not be unique. To see this, consider the situation when $\lambda_q = \lambda_{q+1}$. When this holds, the selection of a q-th eigenvector in $U^\star$ is not unique.

In summary, a solution to problem (6.1) can be obtained as follows. Find the q largest eigenvalues of XX^T and a corresponding set of orthonormal eigenvectors $U^\star$. Then over all q dimensional subspaces, $\mathcal{U}^\star = \mathcal{R}(U^\star)$ minimizes the sum of the squared norms of the projection residuals. By projecting each x_j to $\hat{x}_j = U^\star (U^\star)^T x_j$ we obtain an approximation of the data as points in the subspace $\mathcal{U}^\star$. However, we can also represent $\hat{x}_j$ by its coordinates $c_j = (U^\star)^T x_j$ with respect the orthonormal basis $U^\star$. The vectors $c_j \in \mathbb{R}^q$ uniquely specify the corresponding points $\hat{x}_j$. In doing this transformation we have linearly mapped the data into q-dimensional space.

6.3 An Alternative Viewpoint

We now consider an alternative way to view the same problem. This will give some additional insights into the solution we have derived.

6.3.1 The Sample Covariance of the Data

The data points x_j are "spread out" around the sample mean μ. In the case of scalars, to measure the spread around μ we form the *sample variance* $1/p \sum_{j=1}^{p}(x_j - \mu)^2$. However, for vectors the situation is more complicated since variation about the mean can also depend on direction.

We will continue with our assumption that the data has zero sample mean. Hence we examine how the data is spread around the vector $\mathbf{0}$. Select a unit norm vector $u \in \mathbb{R}^n$ and project x_j onto the line through $\mathbf{0}$ in the direction u. This yields $\hat{x}_j = uu^T x_j$, $j = 1, \ldots, p$. Since the direction is fixed to be u, the projected data is effectively specified by the set of scalars $u^T x_j$. This set of scalars also has zero sample mean:

$$\sum_{j=1}^{p} u^T x_j = u^T \sum_{j=1}^{p} x_j = 0.$$

So the spread of the data in direction u can be quantified by the scalar sample variance

$$\sigma^2(u) = \frac{1}{p}\sum_{j=1}^{n}(u^T x_j)^2 = \frac{1}{p}\sum_{j=1}^{p}(u^T x_j)(u^T x_j)^T = u^T \left(\frac{1}{p}\sum_{j=1}^{p} x_j x_j^T\right) u. \tag{6.3}$$

This expresses the variance of the data as a function of the direction u.

The matrix

$$R = \frac{1}{p}\sum_{j=1}^{p} x_j x_j^T. \tag{6.4}$$

is called the *sample covariance matrix* of the (centered) data. The product $x_j x_j^T$ is a real $n \times n$ symmetric matrix formed by the outer product of the j-th data point with itself. The sample covariance is the mean of these matrices and hence is also a real $n \times n$ symmetric matrix. More generally, if the data is not centered but has sample mean μ, then the sample covariance is

$$R = \frac{1}{p}\sum_{j=1}^{p}(x_j - \mu)(x_j - \mu)^T. \tag{6.5}$$

Lemma 6.3.1. The sample covariance matrix R is symmetric positive semidefinite.

Proof. R is clearly symmetric. Positive semidefiniteness follows by noting that for any $x \in \mathbb{R}^n$,

$$x^T R x = x^T\left(\frac{1}{p}\sum_{j=1}^{p} x_j x_j^T\right)x = \frac{1}{p}\sum_{j=1}^{p}(x^T x_j x_j^T x) = \frac{1}{p}\sum_{j=1}^{p}(x_j^T x)^2 \geq 0.$$

□

6.3.2 Directions of Maximum Variance

Using R and (6.3) we can concisely express the variance of the data in direction u as

$$\sigma^2(u) = u^T R u. \tag{6.6}$$

Hence the direction u in which the data has maximum sample variance is given by the solution of the problem:

$$\begin{aligned} \arg\max_{u \in \mathbb{R}^n} \quad & u^T R u \\ \text{s.t.} \quad & u^T u = 1 \end{aligned} \tag{6.7}$$

with R a symmetric positive semidefinite matrix. This is problem (5.9). By Theorem 5.2.2, the data has maximum variance σ_1^2 in the direction v_1, where $\sigma_1^2 \geq 0$ is the largest eigenvalue of R and v_1 is a corresponding unit norm eigenvector of R.

We must take care if we want to find two directions with the largest variance. Without any constraint, the second direction can be arbitrarily close to v_1 and variance σ_1^2. One way to prevent this is to constrain the second direction to be orthogonal to the first. Then if we want a third direction, constraint it to orthogonal to the two previous directions, and so on. In this case, for q orthogonal directions we want to find $U = [u_1, \ldots, u_q] \in \mathcal{O}_{n,q}$ to maximize $\sum_{j=1}^q u_j^T R u_j = \text{trace}(U^T R U)$. Hence we want to solve problem (6.2) with $P = R$. By Theorem 6.2.1, the solution is attained by taking the q directions to be unit norm eigenvectors $v_1, \ldots, v_q$ for the largest q eigenvalues of R.

By this means you see that we obtain n orthonormal directions of maximum (sample) variance in the data. These directions $v_1, v_2, \ldots, v_n$ and the corresponding variances $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2$ are eigenvectors and corresponding eigenvalues of R: $Rv_j = \sigma_j^2 v_j$, $j = 1, \ldots, n$. The vectors v_j are called the *principal components* of the data, and this decomposition is called *Principal Component Analysis* (PCA). Let V be the matrix with the v_j as its columns, and $\Sigma^2 = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ (note $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2$) . Then PCA is an ordered eigen-decomposition of the sample covariance matrix: $R = V \Sigma^2 V^T$.

There is a clear connection between PCA and finding a subspace that minimizes the sum of squared norms of the residuals. We can see this by writing

$$R = \frac{1}{p} \sum_{j=1}^p x_j x_j^T = \frac{1}{p} X X^T.$$

So the sample covariance is just a scalar multiple of the matrix XX^T. This means that the principal components are just the eigenvectors of XX^T listed in order of decreasing eigenvalues. In particular, the first q principal components are the first q eigenvectors (ordered by eigenvalue) of XX^T. This is exactly the orthonormal basis $U^\star$ that defines an optimal q-dimensional projection subspace $\mathcal{U}^\star$. So the leading q principal components give a particular orthonormal basis for an optimal q-dimensional projection subspace.

A direction in which the data has small variance relative to σ_1^2 may not be not an important direction; after all the data stays close to the mean in this direction. If one accepts this hypothesis, then the directions of largest variance are the important directions. These capture most of the variability in the data. This suggests that we could select an integer $q < \text{rank}(R)$ and project the data onto the q directions of largest variance. Let $V_q = [v_1, v_2, \ldots, v_q]$. Then the projection onto the span of the columns of V_q is $\hat{x}_j = V_q(V_q^T x_j)$. The term $c_j = (V_q^T) x_j$ gives the coordinates of x_j with respect to V_q. Then the product $V_q c_j$ synthesizes $\hat{x}_j$ using these coefficients to form the appropriate linear combination of the columns of V_q.

Here is a critical observation: since the directions are fixed and known, we don't need to form $\hat{x}_j$. Instead we can simply map x_j to the coordinate vector $c_j \in \mathbb{R}^q$. No information is lost in working with c_j instead of $\hat{x}_j$ since the latter is an invertible linear function of the former. Hence $\{c_j\}_{j=1}^p$ gives a new set of data that captures most of the variation in the original data, and lies in a reduced dimension space ($q \leq \text{rank}(R) \leq n$).

The natural next question is how does one select the dimension q? Clearly this involves a tradeoff between the size of q and the amount of variation in the original data that is captured in the projection. The "variance" captured by the projection is $\nu^2 = \sum_{j=1}^q \sigma_j^2$ and the "variance" in the residual is $\rho^2 = \sum_{j=q+1}^n \sigma_j^2$. Reducing q reduces ν^2 and increases ρ^2. The selection of q thus involves determining how much of the total variance in X needs to be captured in order to successfully use the projected data to complete the analysis or decision task at hand. For example, if the projected data is to be used to learn a classifier, then one needs to select the value of q that yields acceptable (or perhaps best) classifier performance. This could be done using cross-validation.

6.4 PCA Computation

We have shown that the q-dimensional PCA projection subspace is spanned by the leading eigenvectors (largest eigenvalues) of XX^T or equivalently of $R = {}^1\!/\!_p XX^T$. If practice it us usually more efficient to compute a compact SVD of X and use this to find the principal components. To see this let $X = U\Sigma V^T$ be a compact SVD. Then

$$XX^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T.$$

Hence the principal components with nonzero variances are the left singular vectors U in a compact SVD of X, and the variance of the data in direction u_j is the square the singular value σ_j.

6.5 Exercises

Exercise 6.1. Let $X \in \mathbb{R}^{n\times p}$. Show that the set of nonzero eigenvalues of XX^T is the same as the set of nonzero eigenvalues of X^TX.

Exercise 6.2. The Rayleigh quotient of matrix $P \in \mathbb{R}^{n\times n}$ evaluated at a nonzero $x \in \mathbb{R}^n$ is

$$R_P(x) = \frac{x^TPx}{x^Tx}.$$

If P is symmetric, show that $\lambda_{\min}(P) \le R_P(x) \le \lambda_{\max}(P)$.

Bibliography

[1] Roger Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1991.