Singular value decomposition, pseudo-inverses, and principal component analysis

Kristian Wichmann

July 8, 2017

1 Gramian matrices

Given a set of vectors $a_1, a_2, \ldots, a_n \in \mathbb{R}^m$, the Gramian matrix is the traditionally matrix of inner products $\langle a_i, a_j \rangle$. If these vectors are collected into a $m \times n$ matrix A, this matrix can be expressed as $A^t A$. Here, we will use the term for any matrix in this form. By starting out with the transpose instead, this means that AA^t is also a Gramian, with dual results.

Theorem 1.1. *If $A \in \mathbb{R}^{m \times n}$, then $A^t A$ is symmetric and positive semi-definite. Iff A has rank m, $A^t A$ is positive definite.*

Proof. $(A^t A)^t = A^t (A^t)^t = A^t A$ shows symmetry. positive semi-definiteness, let $x \in \mathbb{R}^n$. Then:

$$x^t A^t A x = \langle Ax, Ax \rangle = ||Ax||^2 \tag{1.1}$$

As a norm, this is greater than or equal to zero. Hence $A^t A$ is positive semi-definite. If A has rank m the map $x \mapsto Ax$ has a trivial kernel by the rank-kernel theorem. Which means only the zero vector is mapped to zero, and hence $A^t A$ is positive definite. If the rank is less than m, the kernel is non-trivial and positive definiteness cannot be true. □

2 The rank-nullity theorem

2.1 For A and A^t

According to the rank-nullity theorem, for a matrix $A \in \mathbb{R}^{m \times n}$, the sum of the rank and nullity is n. So, if the rank of A is r, then null$A = n - r$. Applying the theorem to A^t, which also has rank r, we get null$A = m - r$.

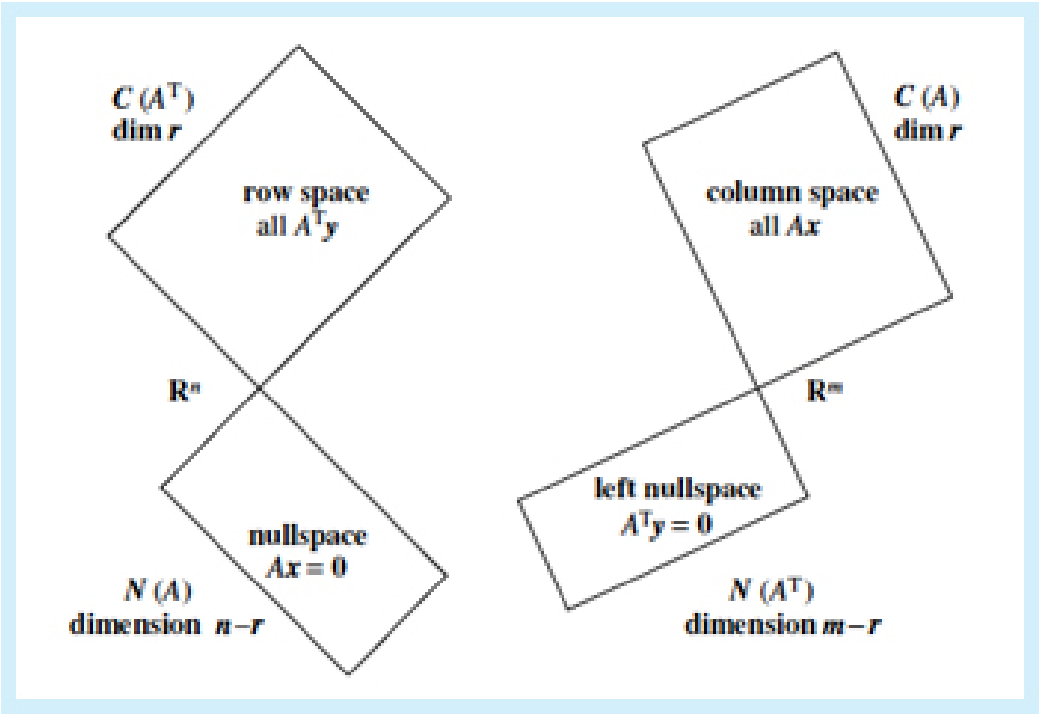

Figure 1: Visualization of dimensionality for the rank-nullity theorem

The image of A is also called the *column space* of A, denoted $C(A)$. The image of A^t is also called the *row space* of A, $C(A^t)$. The null space of A^t is often called the *left null space*.

These relationships are visualized in figure 1.

3 Singular value decomposition

3.1 Construction and intuition

We know that the dimensions of the row and column spaces of a matrix $A \in \mathbb{R}^{m \times n}$ are the same, r. We now seek out orthonormal bases of each of these spaces - $u_1, u_2, \ldots, u_r$ for column space and $v_1, v_2, \ldots, v_r$ for row space, such that

$$Av_i = \sigma_i u_i \tag{3.1}$$

The sigmas are known as *singular values* for A. Now, expand the orthonormal bases to include the null spaces. This means that $Av_i = 0$ for $r < i \leq n$. In matrix form this means:

$$AV = U\Sigma \tag{3.2}$$

Here, the columns of U and V are made from the respective bases, so $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$, and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal $n \times n$ matrix with the σ_i's in the first r places of the diagonal and zeroes in the rest. Solving for A we get:

$$A = U\Sigma V^t \tag{3.3}$$

Here we have used that orthogonal matrices are invertible with their transpose as the inverse. This is the famous *singular value decomposition* of A.

3.2 Finding U and V

The question is how to find U and V? To do so, consider the Gramian matrix of A:

$$A^tA = (U\Sigma V^t)^tU\Sigma V^t = V\Sigma^tU^tU\Sigma V^t = V(\Sigma^t\Sigma)V^t \tag{3.4}$$

But since Σ is diagonal, $(\Sigma^t\Sigma)$ is simply a square, diagonal matrix with $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ in the first r entries of the diagonal and zeroes for the rest. We know that A^tA is symmetric and hence diagonalizable. It is also positive semidefinite and so has non-negative eigenvalues. So we can find use normalized eigenvectors as columns of V and determine the singular values as the square roots of the non-zero eigenvalues.

Similarly, consider AA^t:

$$AA^t = U\Sigma V^t(U\Sigma V^t)^t = U\Sigma V^tV\Sigma^tU^t = U(\Sigma\Sigma^t)U^t \tag{3.5}$$

This is also symmetric and positive semi-definite. Again, $\Sigma\Sigma^t$ is square, this time $m \times m$. It still has the squares of singular values in the diagonal and zeroes for the rest. Now normalized eigenvectors can be used as columns of U.

4 Orthogonal projection

Let U be a subspace of $\mathbb{R}^n$ spanned by the linearly independent set of vectors $a_1, a_2, \ldots, a_m$. Given a $x \in \mathbb{R}^n$, we wish to find a vector u in U, such that $e = x - u$ is orthogonal to U. That means it should be orthogonal to all a_i's:

$$\forall i : \ a_i^t(x - u) = 0 \tag{4.1}$$

This can be expressed in matrix form by collecting all the a_i's into a $n \times m$ matrix A:

$$A = \begin{pmatrix} | & | & \cdots & | \\ a_1 & a_2 & \cdots & a_m \\ | & | & \cdots & | \end{pmatrix} \tag{4.2}$$

Then we may write:

$$A^t(x-u) = 0 \tag{4.3}$$

Since $u \in U$, it can be written as a linear combination of a_i's, so $u = A\beta$. We want to solve for the coefficient vector β:

$$A^t(x - A\beta) = 0 \Leftrightarrow A^t x = A^t A\beta \tag{4.4}$$

Since the a_i's are linearly independent, A^tA is invertible, so:

$$\beta = (A^tA)^{-1}A^t x \tag{4.5}$$

The actual vector is then $A\beta = A(A^tA)^{-1}A^tx$. Which means that the projection operator $p_U : \mathbb{R}^n \to U$ is linear with the corresponding matrix being $P_U = A(A^tA)^{-1}A^t$.

Note, that if the basis spanning U is chosen to be orthonormal, $A^tA = I_m$, and so $(A^tA)^{-1} = I_m$. Hence P_U simplifies to AA^t in this case.

Theorem 4.1. *The matrix P_U is symmetric and idempotent.*

Proof. Both follow directly from the formula $P_U = A(A^tA)^{-1}A^t$:

- Symmetry: $P_U^t = \left(A(A^tA)^{-1}A^t\right)^t = A\left[(A^tA)^{-1}\right]^t A^t$. But since the transpose of an inverse is the inverse of a transpose, and A^tA is symmetric by theorem 1.1 we have $\left[(A^tA)^{-1}\right]^t = \left[(A^tA)^t\right]^{-1} = (A^tA)^{-1}$. Hence $P_U^t = A(A^tA)^{-1}A^t = P_U$.
- Idempotency: $P_U^2 = \left(A(A^tA)^{-1}A^t\right)^2 = A(A^tA)^{-1}A^tA(A^tA)^{-1}A^t = A(A^tA)^{-1}A^t = P_U$.

□

4.1 Projection onto an affine subspace

Sometimes, we may wish to project onto a subspace that does not pass through the origin, but rather is simply parallel to a proper subspace U. Such a space is known as an *affine subspace*, and can be written as:

$$W = w + U = \{w + u | u \in U\} \tag{4.6}$$

Here w is any vector in the affine space (and thus the choice of w is not unique, so $v + S = w + S$ does not imply that $v = w$). We can find a formula for projection unto W by translating to a coordinate system where W passes through the origin. This is done by subtracting w. The formula above can

then be applied to project onto U, and finally we need to translate back into the origin coordinate system by adding w. To sum it up:

$$P_W x = w + P_U(x - w) \tag{4.7}$$

Above P_W and P_U denotes projection onto W and U respectively. If U is spanned by linearly independent columns of the matrix A, then we know from above that[1]

$$P_U = (A^t A)^{-1} A^t \tag{4.8}$$

Now equation 4.7 turns into:

$$P_W x = w + (A^t A)^{-1} A^t (x - w) \tag{4.9}$$

If $x = 0$ is the point being projected we get:

$$P_W 0 = w - (A^t A)^{-1} A^t w) = (I - (A^t A)^{-1} A^t) w \tag{4.10}$$

5 Generalized inverses

For an invertible matrix A, it's obviously true that:

$$AA^{-1}A = A \tag{5.1}$$

If A is not invertible, we may still define a *generalized inverse* A^g as a matrix that satisfies the same equation:

$$AA^g A = A \tag{5.2}$$

If A^g further satisfies:

$$A^g A A^g = A^g, \tag{5.3}$$

it is called a *reflexive generalized inverse*.

5.1 Left inverses

If $A \in \mathbb{R}^{m \times n}$ has rank n, then the null space is trivial, and hence the corresponding linear transformation is injective. This means that the equation $Ax = b$ may or may not have a solution, but if it exists, it's unique. In particular, if $n = m$ the existence is guaranteed, but if $n < m$ it's possible,

[1] This assumes that we're dealing with coordinate space.

but unlikely. The matrix A^tA has rank n as well, and hence is invertible. This can be used to construct a left inverse:

$$A_L^{-1} = (A^tA)^{-1}A^t, \qquad A_L^{-1}A = (A^tA)^{-1}A^tA = I_n \tag{5.4}$$

But we already know from the last section that A_L^{-1} is more or less the projection operator unto the image space of A: In fact it's the coordinate vector with respect to the basis of column vectors of A. This means that $AA_L^{-1}b$ is the vector in the image space that is closest to b.

5.1.1 Example

Consider the equation:

$$\begin{pmatrix}3\\4\end{pmatrix} x = \begin{pmatrix}7\\1\end{pmatrix} \tag{5.5}$$

Here x is a 1 by 1 matrix (or simply a real number). It is immediately clear, that this equation has no solutions. The situation is visualized in figure 2: The point $\begin{pmatrix}7\\1\end{pmatrix}$ clearly does not lie on the line traced by $\begin{pmatrix}3\\4\end{pmatrix}$

Using the general notation, here $A = \begin{pmatrix}3\\4\end{pmatrix} \in \mathbb{R}^{2\times 1}$ has rank 1, and so a left inverse can be found:

$$A_L^{-1} = (A^tA)^{-1}A^t = \left(\begin{pmatrix}3 & 4\end{pmatrix}\begin{pmatrix}3\\4\end{pmatrix}\right)^{-1}\begin{pmatrix}3 & 4\end{pmatrix} = \frac{1}{25}\begin{pmatrix}3 & 4\end{pmatrix} \tag{5.6}$$

The best approximation to a solution is then:

$$x = A_L^{-1}b = \frac{1}{25}\begin{pmatrix}3 & 4\end{pmatrix}\begin{pmatrix}7\\1\end{pmatrix} = \frac{21+4}{25} = 1 \tag{5.7}$$

So, the actual point in the image space is:

$$Ax = \begin{pmatrix}3\\4\end{pmatrix} 1 = \begin{pmatrix}3\\4\end{pmatrix} \tag{5.8}$$

5.2 Right inverses

Similarly, if $A \in \mathbb{R}^{m\times n}$ has rank m, then the image space is all of $\mathbb{R}^m$, and hence the corresponding linear transformation is surjective. This means that the equation $Ax = b$ always has a solution, and it will have infinitely many if $n > m$ (as we will see below). The matrix AA^t has rank m as well, and hence is invertible. Analogously, we can use this to construct a right inverse:

$$A_R^{-1} = A^t(AA^t)^{-1}, \qquad AA_R^{-1} = AA^t(AA^t)^{-1} = I_m \tag{5.9}$$

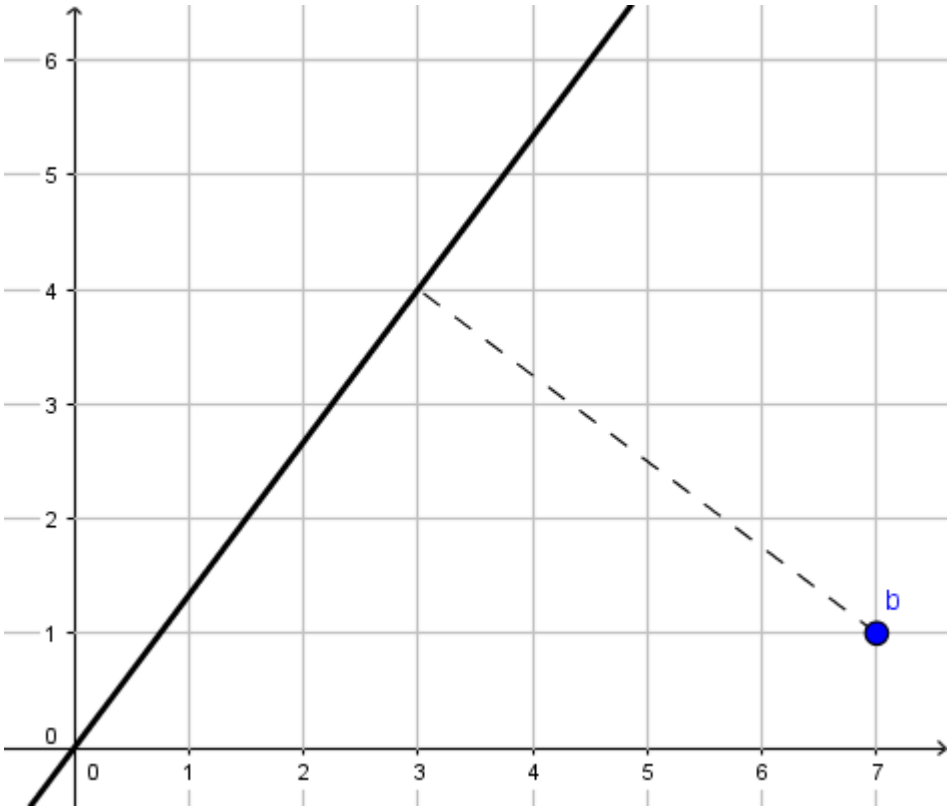

Figure 2: The geometry of equation 5.5

This means that $x_s = A_R^{-1}b$ is a solution.

To get an intuition for which x is picked out, consider instead the equation $Ax = 0$. The solution space S is the null space of A. According to the rank-nullity theorem S has dimension $s = m - n$. Let $l_1, l_2, \cdots, l_s$ be a basis for S. These are collected into the $s \times n$ matrix L:

$$L = \begin{pmatrix} | & | & \cdots & | \\ l_1 & l_2 & \cdots & l_s \\ | & | & \cdots & | \end{pmatrix} \tag{5.10}$$

Now, since we have one solution to the equation $Ax = b$, namely x_s, it follows that any vector of the form $s_x + s$, where $s \in S$ is a solution. But this is exactly the affine space $W = x_s + S$ as described in the section on projection above. We now ask: What is the projection of the zero vector onto W? According to equation 4.10 it is:

$$P_W 0 = (I - (L^t L)^{-1} L^t) w \tag{5.11}$$

Here w is a vector in W. We choose $w = x_s A_R^{-1} b = A^t (AA^t)^{-1} b$:

$$P_W 0 = (I - (L^t L)^{-1} L^t) A^t (AA^t)^{-1} b \tag{5.12}$$

But since $L^t A^t = (AL)^t = 0$ because of L's construction, the giant cross-term between A's and L's vanishes, and we're left with:

$$P_W 0 = A^t (AA^t)^{-1} b = x_s \tag{5.13}$$

In other words, the solution obtained by using A_R^{-1} is the projection of the zero vector onto the solution space. This can be characterized as the solution vector with the smallest possible norm.

5.2.1 Example

Consider the equation:

$$\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 4 \tag{5.14}$$

This clearly has infinitely many solutions, as it is equivalent to $x + y = 4$ or $y = -x + 4$. Let's consider the right inverse of $A = \begin{pmatrix} 1 & 1 \end{pmatrix}$:

$$A_R^{-1} = A^t (AA^t)^{-1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left(\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)^{-1} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{5.15}$$

Hence, the solution picked out by A_R^{-1} is:

$$\begin{pmatrix} x \\ y \end{pmatrix} = A_R^{-1} b = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} 4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \tag{5.16}$$

The situation is shown in figure 3.

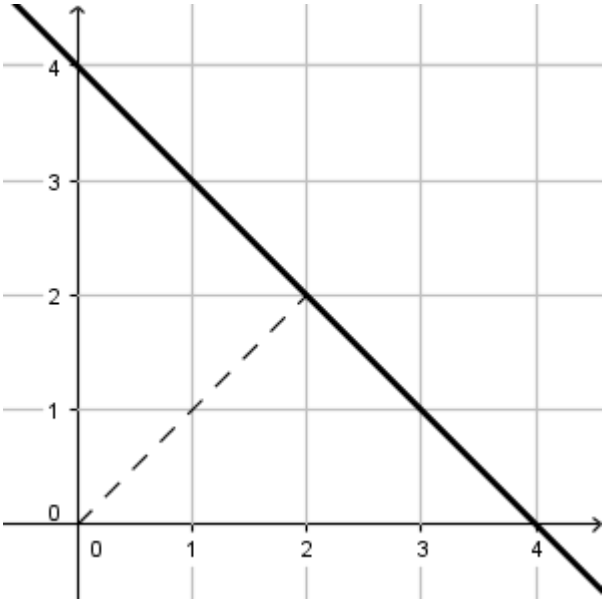

Figure 3: The geometry of equation 5.14

5.3 Reflexive, generalized inverses

Both of of these inverses (when they exist) satisfies equation 5.2. They also satisfy 5.3. For instance:

$$A_L^{-1}AA_L^{-1} = (A^tA)^{-1}A^tA(A^tA)^{-1}A^t = (A^tA)^{-1}A^t = A_L^{-1} \tag{5.17}$$

So both are reflexive, generalized inverses.

6 The Moore-Penrose pseudoinverse

The *Moore-Penrose pseudoinverse* or simply the pseudoinverse of a real matrix A is the reflexive, generalized inverse A^+ which also satisfies:

$$(AA^+)^t = AA^+, \qquad (A^+A)^t = A^+A \tag{6.1}$$

In other words, for which AA^+ and A^+A are symmetrical[2]. Summing it all up, the pseudoinverse has to satisfy the following four conditions:

1. $AA^+A = A$
2. $A^+AA^+ = A^+$

[2]If A has complex entries, the condition instead becomes that A should be Hermitian.

3. $(AA^+)^t = AA^+$

4. $(A^+A)^t = A^+A$

6.1 Uniqueness

If such a pseudoinverse exists, it is unique (hence our use of definite article above). To show this, let B_1 and B_2 be pseudoinverses of A. Then:

$$AB_1 = (AB_1)^t = B_1^t A^t = B_1^t (AB_2 A)^t = B_1^t A^t B_2^t A^t = \tag{6.2}$$

$$(AB_1)^t (AB_2)^t = AB_1 AB_2 = AB_2 \tag{6.3}$$

Similarly:

$$B_1 A = (B_1 A)^t = A^t B_1^t = (AB_2 A)^t B_1^t = A^t B_2^t A^t B_1^t = \tag{6.4}$$

$$(B_2 A)^t (B_1 A)^t = B_2 A B_1 A = B_2 A \tag{6.5}$$

But then:

$$B_1 = B_1 A B_1 = B_2 A B_1 = B_2 A B_2 = B_2 \tag{6.6}$$

6.2 Intuition behind the pseudoinverse

The idea behind the pseudoinverse is similar to the one used in singular value decomposition: The dimension of the column and row spaces of a matrix $A \in \mathbb{R}^{m \times n}$ have the same dimension, r. So if $y \in \mathbb{R}^m$ is in the column space, there is exactly one vector $x \in \mathbb{R}^n$ so that $Ax = y$. However, for y in the left null space, we're in trouble. But what if we just send these these vectors to the zero vector? This corresponds to projecting onto the column space.

6.3 Definition for diagonal matrices

Let D be a diagonal, but not necessarily square matrix, i.e. $D \in \mathbb{R}^{m \times n}$. The diagonal entries are denoted d_i. So $D_{ij} = d_i \delta_{ij}$. Following the intuition section, we will set the pseudoinverse of D to be the diagonal $n \times m$ matrix whose diagonal entries are set equal to:

$$d_i^+ = \begin{cases} 0 \text{ for } d_i = 0 \\ \frac{1}{d_i} \text{ otherwise} \end{cases} \tag{6.7}$$

We can use this to write the entries of D^+ as $D_{ij}^+ = (D_{ji})^+ = d_i^+ \delta_{ij}$.

6.3.1 Checking the axioms

We must now check that the four points above are satisfied:

1. Calculate the indices of DD^+D:

$$(DD^+D)_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{m} D_{ik}D^+_{kl}D_{lj} = \sum_{k=1}^{n}\sum_{l=1}^{m} d_i\delta_{ik}d^+_k\delta_{kl}d_l\delta_{lj} \tag{6.8}$$

   Two of the deltas cancel out, so:

$$(DD^+D)_{ij} = d_i d^+_i d_i\delta_{ij} \tag{6.9}$$

   Now, consider the product of the first two terms:

$$d_i d^+_i = \begin{cases} 0 \text{ if } d_i = 0 \\ 1 \text{ if } d_i \neq 0 \end{cases} \tag{6.10}$$

   Multiplying by d_i we get d_i in both cases, and so:

$$(DD^+D)_{ij} = d_i\delta_{ij} \tag{6.11}$$

2. Using the same strategy:

$$(D^+DD^+)_{ij} = \sum_{k=1}^{m}\sum_{l=1}^{m} D^+_{ik}D_{kl}D^+_{lj} = \sum_{k=1}^{n}\sum_{l=1}^{m} d^+_i\delta_{ik}d_k\delta_{kl}d^+_l\delta_{lj} \tag{6.12}$$

   Again, two deltas cancel:

$$(D^+DD^+)_{ij} = d^+_i d_i d^+_i\delta_{ij} \tag{6.13}$$

   By using equation 6.10 again we get:

$$(D^+DD^+)_{ij} = d^+_i\delta_{ij} \tag{6.14}$$

3. Same strategy:

$$(DD^+)^t_{ij} = (DD^+)_{ji} = \sum_{k=1}^{n} D_{jk}D^+_{ki} = \sum_{k=1}^{n} d_j\delta_{jk}d^+_k\delta_{ki} = d_i d^+_i\delta_{ij} \quad = \tag{6.15}$$

$$d^+_i d_i\delta ij = \sum_{k=1}^{m} d^+_i\delta_{ik}d_k\delta_{kj} = \sum_{k=1}^{m} D_{ik}D^+_{kj} = (DD^+)_{ij} \tag{6.16}$$

4. Analogous to 3.

6.4 Definition for arbitrary matrices

We can now use the singular value decomposition to generalize the pseudoinverse. If $A \in \mathbb{R}^{m \times n}$ is an arbitrary matrix, then we may write $A = UDV^t$ as above. Now, we set the psuedoinverse to:

$$A^+ = VD^+U^t \tag{6.17}$$

6.5 Checking the axioms

We can now show that this definition obeys the four axioms above. Each of the proofs hinge on the result for diagonal matrices:

1. $AA^+A = UDV^tVD^+U^tUDV^t = UDD^+DV^t = UDV^t = A$
2. $A^+AA^+ = VD^+U^tUDV^tVD^+U^t = VD^+DD^+U^t = VD^+U^t = A^+$
3. $(AA^+)^t = (UDV^tVD^+U^t)^t = (UDD^+U^t) =$
   $U(DD^+)^tU^t = UDD^+U^t = UDV^tVD^+U^t = AA^+$
4. $(A^+A)^t = (VD^+U^tUDV^t)^t = (VD^+DV^t) =$
   $V(D^+D)^tV^t = VD^+DV^t = VD^+U^tUDV^t = A^+A$

6.6 Left and right inverses as pseudoinverses

In the previous section, we looked at left and right inverses as projection operators used for getting the best solution to linear equation systems. It turns out, that these are in fact the pseudoinverse of A in their respective cases. This is done by checking that they satisfy the four axioms. The result then follows by the uniqueness of the pseudoinverse.

6.6.1 Left inverses

As above, for A injective, we set $A_L = (A^tA)^{-1}A^t$. We need to show the four axioms are satisfied:

- $AA_LA = A(A^tA)^{-1}A^tA = A$.
- $A_LAA_L = (A^tA)^{-1}A^tA(A^tA)^{-1}A^t = (A^tA)^{-1}A^t = A_L$.
- $(AA_L)^t = (A(A^tA)^{-1}A^t)^t = A(A^tA)^{-1}A^t = AA_L$.
- $(A_LA)^t = I^t = I = A_LA$.

6.6.2 Right inverses

Similarly, when A is surjective, we set $A_R = A^t(AA^t)^{-1}$. As above, we need to show it satisfies the four axioms:

- $AA_RA = AA^t(AA^t)^{-1}A = A$.
- $A_RAA_R = A^t(AA^t)^{-1}AA^t(AA^t)^{-1} = A^t(AA^t)^{-1} = A_R$.
- $(AA_R)^t = (AA^t(AA^t)^{-1})^t = I^t = I = AA_R$.
- $(A_RA)^t = (A^t(AA^t)^{-1}A)^t = A^t(AA^t)^{-1}A = A_RA$.

6.7 Pseudoinverses as limits

The pseudoinverse can also be found through limit procedures:

$$A^+ = \lim_{\delta \to 0^+} (A^tA + \delta I)^{-1}A^t = \lim_{\delta \to 0^+} A^t(AA^t + \delta I)^{-1} \tag{6.18}$$

To see why this is true, consider the singular value decomposition for A and A^t:

$$A = U\Sigma V^t, \quad A^t = V\Sigma U^t \tag{6.19}$$

This is not strictly true: In general, the sigma matrices will have differing sizes, but since the only true difference is that one has more zeroes in the diagonal than the other, we will slightly abuse notation and use the same symbol for each. Now the above means that:

$$AA^t = U\Sigma V^tV\Sigma U^t = U\Sigma U^t, \quad A^tA = V\Sigma U^tU\Sigma V^t = V\Sigma^2V^t \tag{6.20}$$

Now consider the matrix to be inverted in the first limit formula:

$$A^tA + \delta I = V\Sigma^2V^t + \delta VV^t = V(\Sigma^2 + \delta)V^t \tag{6.21}$$

Hence, the eigenvalues of the matrix are:

$$\lambda_i = \frac{1}{\sigma_i^2 + \delta} \tag{6.22}$$

Since σ_i^2 is non-negative, and δ is positive, this means that all eigenvalues are positive, and hence the inverse exists. Collecting all these eigenvalues into a diagonal matrix Λ this means that the inverse can be written $V\Lambda^{-1}V^t$. Now multiply by A^t:

$$(A^tA + \delta I)^{-1}A^t = V\Lambda^{-1}V^tV\Sigma U^t = V\Lambda^{-1}\Sigma U^t \tag{6.23}$$

This is a singular value decomposition, with singular values:

$$\frac{\sigma_i}{\sigma_i^2 + \delta} \tag{6.24}$$

As $\delta \to 0^+$ we get $1/\sigma_i$, unless $\sigma_i = 0$, then we get zero. This means that we get exactly what the recipe for constructing the pseudoinverse said!

The second limit formula follows from a completely analogous argument.

7 Principal component analysis

Principal component analysis (or simply PCA) is an algorithm that achieves lossy compression of the information in a data set using SVD.

7.1 Singular value decomposition of data

Let X be an $n \times p$ matrix containing p data points of dimensionality n. We will assume that the data has been *normalized*, so that the average of each of the n dimensions is zero. In this case, we may write the empirical covariance matrix as:

$$C = \frac{X^t X}{n-1} \tag{7.1}$$

But, we may also do a singular value decomposition of X:

$$X = USV^t \tag{7.2}$$

Inserting equation 7.2 into 7.1 we get:

$$C = \frac{1}{n-1} VSU^t USV^t = \frac{1}{n-1} VS^2V^t \tag{7.3}$$

So, the relationship between the eigenvalues of the covariance matrix and the singular values of S are:

$$\lambda_i = \frac{1}{n-1} s_i^2 \tag{7.4}$$

The orthogonal set of rows vectors of V - the eigenvectors of C - are called *principal axes* and the corresponding *principal components* are the rows of $XV = USV^tV = US$.

7.2 Autoencoders

An *autoencoder* of a set of data points $x_1, \cdots, x_p \in \mathbb{R}^n$ consists of two parts:

- A *coder* function $f : \mathbb{R}^n \to \mathbb{R}^m$, which turns an input vector $x \in \mathbb{R}^n$ into an encoded vector $c = f(x) \in \mathbb{R}^m$.
- A *decoder* function $g : \mathbb{R}^m \to \mathbb{R}^n$ which transforms encoded vectors back into vectors from the original space $\mathbb{R}^n$.

The coder and decoder should ideally be chosen such that for an input x, we get the exact input back when we decode the coded input. In this case the autoencoder is *lossless*:

$$\forall i : \ g(f(x_i)) = x_i \tag{7.5}$$

Here, we will consider *lossy* compression, where this is only approximately true:

$$\forall i : \ g(f(x_i)) \approx x_i \tag{7.6}$$

What this means exactly will vary according to the approach taken.

7.3 Linear decoder

Here, we will require the decoder to be linear. Hence it can be expressed in matrix form:

$$g(c) = Dc, \tag{7.7}$$

where $D \in \mathbb{R}^{n \times m}$. This means that the decoded vectors lie in a subspace of $\mathbb{R}^n$ with a dimension of m or less - in the following we assume this dimension to be m. We can think of the coded vector c as coordinates for the basis specified by the rows of D. There will be several ways to choose the same transformation, so we must expect to restrict the structure of D somehow. More on that below.

Now, consider a coded vector $c = f(x)$. A sensible way to choose D is to require that the squared distance between x and $g(f(x)) = Dc$ to be as small as possible. I.e. we seek to minimize:

$$||Dc - x||^2 = (Dc - x)^t(Dc - x) = (c^t D^t - x^t)(Dc - x) = \tag{7.8}$$

$$c^t D^t Dc - c^t D^t x - x^t Dc + x^t x \tag{7.9}$$

Which encoding minimizes this distance? Let's find out by differentiating with respect to c:

$$\frac{\partial ||Dc - x||^2}{\partial c} = 2D^t Dc - D^t x - D^t x = 2(D^t Dc - D^t x) \tag{7.10}$$

This is zero when:

$$D^tDc - D^tx = 0 \Leftrightarrow c = (D^tD)^{-1}D^tx \tag{7.11}$$

So the optimal encoding in this sense will be linear as well. This means that the entire coding/decoding can be written as:

$$g(f(x)) = D(D^tD)^{-1}D^tx \tag{7.12}$$

I.e. an orthogonal projection on the image space of D.

Now, it is an obvious choice to require the row vectors of D to be orthogonal, as it simplifies the projection operator to DD^t:

$$g(f(x)) = DD^tx \tag{7.13}$$

This constraint can be written as $D^tD = I_m$.

7.4 Full data set

This is just one data point however. We have p such points. Because of the linearity we can conveniently write the encoding of all p points in one matrix equation:

$$g(f(X)) = DD^tX \in \mathbb{R}^{n\times p} \tag{7.14}$$

To get the differences between this and the original points, subtract X:

$$\Delta = X - DD^tX = (I_n - DD^t)X \tag{7.15}$$

We may view this matrix as p row vectors of differences:

$$\Delta = \begin{pmatrix} | & \cdots & | \\ \delta_1 & \cdots & \delta_p \\ | & \cdots & | \end{pmatrix} \tag{7.16}$$

We wish to minimize to total, squared distance. To get this, consider the following product:

$$\Delta^t\Delta = \begin{pmatrix} - & \delta_1^t & - \\ \vdots & \vdots & \vdots \\ - & \delta_p^t & - \end{pmatrix} \begin{pmatrix} | & \cdots & | \\ \delta_1 & \cdots & \delta_p \\ | & \cdots & | \end{pmatrix} \tag{7.17}$$

So each entry is a dot product between differences, $[\Delta^t\Delta]_{ij} = \delta_i^t\delta_j$. However, we're only interested in the cases where $i = j$. Which means we need to sum

over the diagonal: We seek to minimize the *trace* of $\Delta^t\Delta$. However, $\Delta^t\Delta$ can also be written as:

$$X^t(I_n - DD^t)^t(I_n - DD^t)X = X^t(I_n - DD^t)(I_n - DD^t)X = X^t(I_n - DD^t)X \tag{7.18}$$

Here we've used that since DD^t is a projection operator, so is $I_n - DD^t$. Hence it is idempotent. The trace is:

$$\text{tr}[X^t(I_n - DD^t)X] = \text{tr}[X^tX] - \text{tr}[X^tDD^tX] \tag{7.19}$$

The first term is just a constant. This means that the problem reduces to maximizing $\text{tr}[X^tDD^tX]$, still subject to $D^tD = I_m$.

7.5 Connection to covariance

We now wish to show, that this is equivalent to maximizing the trace over the quadratic form D^tCD, where C is the covariance matrix. Note that we're still assuming the data to be normalized, such that equation 7.1 holds. More specifically, we will prove that:

$$\text{tr}[X^tDD^tX] = (n-1)\text{tr}[D^tCD] \tag{7.20}$$

This is done by induction over m.

7.5.1 Base case: $m = 1$

First consider the case where $m = 1$. Here $D = d \in \mathbb{R}^{n\times 1}$. The constraint is $d^td = 1$. We seek to maximize:

$$\text{tr}[X^tdd^tX] = \sum_{i=1}^{p}(X^tdd^tX)_{ii} = \sum_{i=1}^{p}\sum_{j=1}^{n} X_{ij}^t d_i d_j X_{ji} \tag{7.21}$$

Here, we have explicitly written out the sum, and used that d only has one dimension: $d_j = D_{j1}$. Now, since the d's are numbers, these can be commuted to yield:

$$\sum_{i=1}^{p}\sum_{j=1}^{n} d_i X_{ij}^t X_{ji} d_j = \text{tr}[d^tX^tXd] = d^tX^tXd \tag{7.22}$$

In the last step, we've used that the result is 1×1. Now, under the assumption that the data are centered, this is essentially a quadratic form in the covariance matrix of the data. Recall that according to equation 7.1

$$X^tX = (n-1)C \tag{7.23}$$

So:

$$\text{tr}[X^tdd^tX] = (n-1)d^tCd \tag{7.24}$$

Since d^tCd is just a number, it is equal to its own trace. and we're done.

7.5.2 A practical lemma

The following result will be useful, both for the induction step and the maximization itself:

Theorem 7.1. *Let $A \in \mathbb{R}^{n\times m}$ be a matrix split into block form as follows:*

$$A = \begin{pmatrix} A' & | & a \end{pmatrix}, \quad A' \in \mathbb{R}^{n\times(m-1)}, a \in \mathbb{R}^{n\times 1} \tag{7.25}$$

Furthermore, let $B \in \mathbb{R}^{n\times n}$. Then:

$$tr[A^tBA] = tr[(A')^tBA'] + a^tBa \tag{7.26}$$

Proof. Write out the trace:

$$\mathrm{tr}[A^tBA] = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} \begin{pmatrix} (A')^t \\ - \\ a \end{pmatrix}_{ij} B_{jk} \begin{pmatrix} A' & | & a \end{pmatrix}_{ki} \tag{7.27}$$

Now, split the i-sum into two, one from 1 to $m-1$ and the last term where $i = m$.

$$\sum_{i=1}^{m-1}\sum_{j=1}^{n}\sum_{k=1}^{n}(A')^t_{ij}B_{jk}A_{ki} + \sum_{j=1}^{n}\sum_{k=1}^{n}a_jb_{jk}a_k = \mathrm{tr}[(A')^tBA'] + a^tBa \tag{7.28}$$

□

7.5.3 Induction step: $(m-1) \Rightarrow (m)$

Assume the hypothesis true for $m-1$. Now we wish to show that this implies truth for m as well. In the latter case, we may write D in the block form:

$$D = \begin{pmatrix} D' & | & d \end{pmatrix}, \quad D' \in \mathbb{R}^{n\times(m-1)}, d \in \mathbb{R}^{n\times 1} \tag{7.29}$$

Now, use the cyclical property of the trace:

$$\mathrm{tr}[X^tDD^tX] = \mathrm{tr}[D^tXX^tD] \tag{7.30}$$

Now use theorem 7.1 with $A = D$ and $B = XX^t$:

$$\mathrm{tr}[D^tXX^tD] = \mathrm{tr}[(D')^tXX^tD'] + d^tXX^td \tag{7.31}$$

Now use the cyclical property again on the first term, and use the same rearranging in the second term as we did in the base case:

$$\mathrm{tr}[X^tDD^tX] = \mathrm{tr}[(D')^tXX^tD'] + d^tX^tXd \tag{7.32}$$

According to the induction hypothesis, the first term is $(n-1)\mathrm{tr}[(D')^tCD^t]$. And from the induction step reasoning, we know the the second term is $(n-1)d^tCd$. Now we can use theorem 7.1 "in reverse":

$$(n-1)\mathrm{tr}[(D')^tCD^t] + (n-1)d^tCd = (n-1)\mathrm{tr}[D^tCD] \tag{7.33}$$

7.6 Solving the maximization problem

So, having proven that $\text{tr}[X^t DD^t X] = (n-1)\text{tr}[D^t CD]$. For optimization, the constant $n-1$ is irrelevant. Now, use theorem 7.1 repeatedly on $\text{tr}[D^t CD]$ to get:

$$\text{tr}[D^t CD] = d_1^t Cd_1 + \cdots + d_m^t Cd_m \tag{7.34}$$

Still, this is subject to $D^t D = I_m$. To maximize such a sum, we need to maximize each of the summands. So we have a set of maximization problems with a constraint on our hands. Such a constraint can be dealt with through a Lagrange multiplier. The constraint on each row vector is:

$$d_i^t d_i = 1 \Leftrightarrow d_i^t d_i - 1 = 0 \tag{7.35}$$

So, the corresponding Lagrangian for the i'th equation is:

$$L = d_i^t Cd_i + \lambda(d_i^t d_i - 1) \tag{7.36}$$

Now find the derivatives:

$$\frac{\partial L}{\partial d_i} = 2Cd_i - 2\lambda d_i \tag{7.37}$$

Setting this equal to zero, we get an eigenvalue problem:

$$Cd_i = \lambda d_i \tag{7.38}$$

So we search for the normalized eigenvector of C which maximizes $d_i^t Cd_i = d_i^t \lambda d_i = \lambda$. So we need to pick the largest eigenvalues and the corresponding eigenvectors. We can only pick each value once[3] because of the requirement that the rows are orthogonal. In short, we need to pick out the m largest eigenvalues of C and have the rows of D be the corresponding (normalized) eigenvectors - the principal axes.

[3]Unless the corresponding eigenspace has more than one dimension.