

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

Kristian Wichmann

May 10, 2017

## 1 Gramian matrices

Given a set of vectors  $a_1, a_2, \dots, a_n \in \mathbb{R}^m$ , 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:

$$x^t A^t A x = \langle Ax, Ax \rangle = \|Ax\|^2 \quad (1.1)$$

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.  $\square$

## 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  $\text{null}A = n - r$ . Applying the theorem to  $A^t$ , which also has rank  $r$ , we get  $\text{null}A = m - r$ .

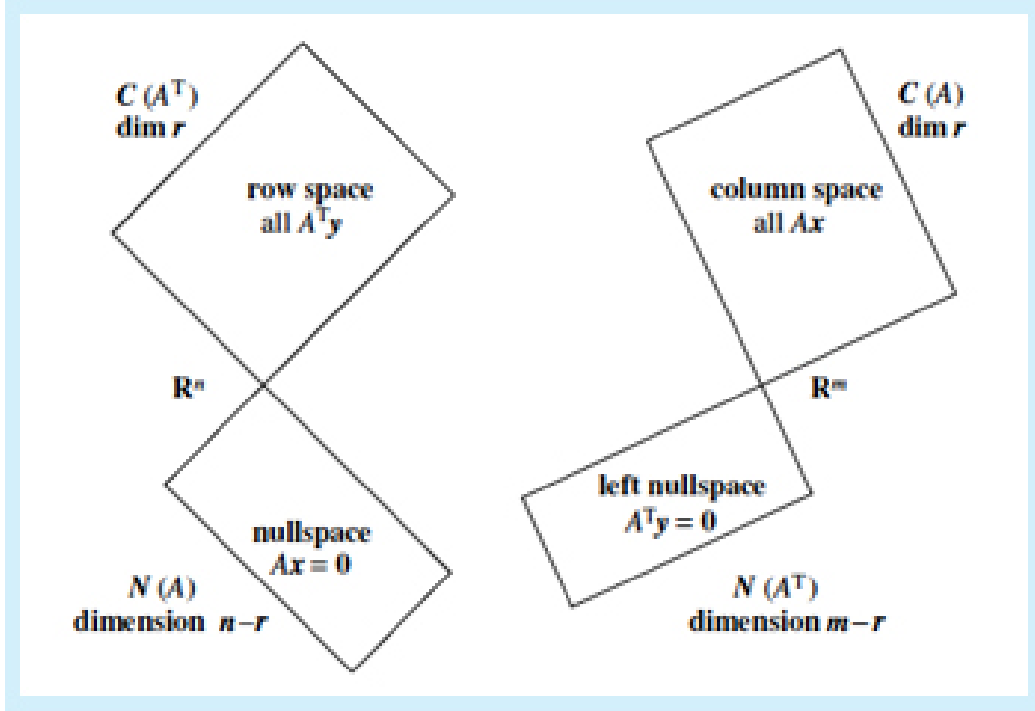


Figure 1: Visualization of dimensionality for the rank-nullity theorem

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 -  $u_1, u_2, \dots, u_r$  for column space and  $v_1, v_2, \dots, v_r$  for row space, such that

$$Av_i = \sigma_i u_i \quad (3.1)$$

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:

$$AV = U\Sigma \quad (3.2)$$

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  $\sigma_i$ 's in the first  $r$  places of the diagonal and zeroes in the rest. Solving for  $A$  we get:

$$A = U\Sigma V^t \quad (3.3)$$

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  $A$ :

$$A^t A = (U\Sigma V^t)^t U\Sigma V^t = V\Sigma^t U^t U\Sigma V^t = V(\Sigma^t \Sigma) V^t \quad (3.4)$$

But since  $\Sigma$  is diagonal,  $(\Sigma^t \Sigma)$  is simply a square, diagonal matrix with  $\sigma_1^2, \sigma_2^2, \dots, \sigma_r^2$  in the first  $r$  entries of the diagonal and zeroes for the rest. We know that  $A^t A$  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  $AA^t$ :

$$AA^t = U\Sigma V^t (U\Sigma V^t)^t = U\Sigma V^t V\Sigma^t U^t = U(\Sigma \Sigma^t) U^t \quad (3.5)$$

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  $a_1, a_2, \dots, a_m$ . 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:

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

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

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

Then we may write:

$$A^t(x - u) = 0 \quad (4.3)$$

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  $\beta$ :

$$A^t(x - A\beta) = 0 \Leftrightarrow A^t x = A^t A \beta \quad (4.4)$$

Since the  $a_i$ 's are linearly independent,  $A^t A$  is invertible, so:

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

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

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

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

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

□

## 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:

$$W = w + U = \{w + u | u \in U\} \quad (4.6)$$

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 onto  $W$  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:

$$P_W x = w + P_U(x - w) \quad (4.7)$$

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<sup>1</sup>

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

Now equation 4.7 turns into:

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

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

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

## 5 Generalized inverses

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

$$A A^{-1} A = A \quad (5.1)$$

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

$$A A^g A = A \quad (5.2)$$

If  $A^g$  further satisfies:

$$A^g A A^g = A^g, \quad (5.3)$$

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

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

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  $A A_L^{-1} b$  is the vector in the image space that is closest to  $b$ .

---

<sup>1</sup>This assumes that we're dealing with coordinate space.

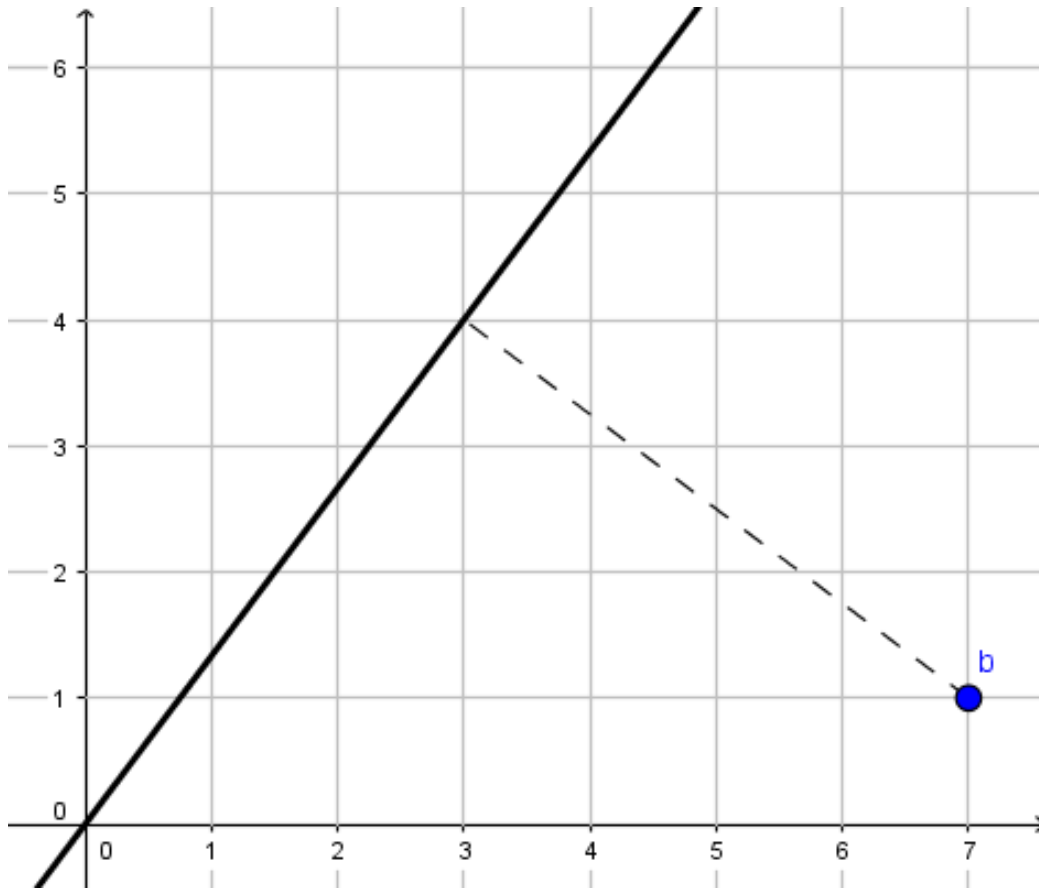


Figure 2: The geometry of equation 5.5

### 5.1.1 Example

Consider the equation:

$$\begin{pmatrix} 3 \\ 4 \end{pmatrix} x = \begin{pmatrix} 7 \\ 1 \end{pmatrix} \quad (5.5)$$

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:

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

The best approximation to a solution is then:

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

So, the actual point in the image space is:

$$Ax = \begin{pmatrix} 3 \\ 4 \end{pmatrix} 1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \quad (5.8)$$

## 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:

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

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  $l_1, l_2, \dots, l_s$  be a basis for  $S$ . These are collected into the  $s \times n$  matrix  $L$ :

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

Now, since we have one solution to the equation  $Ax = b$ , namely  $x_s$ , it follows that any vector of the form  $x_s + s$ , 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:

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

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

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

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:

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

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:

$$(1 \ 1) \begin{pmatrix} x \\ y \end{pmatrix} = 4 \quad (5.14)$$

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  $A = (1 \ 1)$ :

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

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

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

### 5.3 Reflexive, generalized inverses

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

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

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:

$$(AA^+)^t = AA^+, \quad (A^+A)^t = A^+A \quad (6.1)$$

In other words, for which  $AA^+$  and  $A^+A$  are symmetrical.

### 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:

$$AB_1 = (AB_1)^t = B_1^tA^t = B_1^t(AB_2A)^t = B_1^tA^tB_2^tA^t = \quad (6.2)$$

$$(AB_1)^t(AB_2)^t = AB_1AB_2 = AB_2 \quad (6.3)$$



Similarly:

$$B_1 A = (B_1 A)^t = A^t B_1^t = (A B_2 A)^t B_1^t = A^t B_2^t A^t B_1^t = \quad (6.4)$$

$$(B_2 A)^t (B_1 A)^t = B_2 A B_1 A = B_2 A \quad (6.5)$$

But then:

$$B_1 = B_1 A B_1 = B_2 A B_1 = B_2 A B_2 = B_2 \quad (6.6)$$

## 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.

## 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 Autoencoders

An *autoencoder* of a set of data points  $x_1, \dots, x_m \in X = \mathbb{R}^n$  consists of two parts:

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

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*:

$$\forall i : g(f(x_i)) = x_i \quad (7.1)$$

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

$$\forall i : g(f(x_i)) \approx x_i \quad (7.2)$$

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

## 7.2 Linear, "orthogonal" decoder

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

$$g(c) = Dc, \tag{7.3}$$

where  $D \in \mathbb{R}^{n \times l}$ . Furthermore, we will require the columns of  $D$  to be orthonormal to each other. This is of course only possible if  $l \leq n$ , but otherwise there would be no compression, so this is entirely reasonable.