Singular Value Decomposition

Ryan McCorvie[*]

October 16^{th}, 2019

1 The Singular Value Decomposition

I am very indebted to Jean Gallier and Jocelyn Quaintance online notes "Algebra, Topology, Differential Calculus, and Optimization Theory For Computer Science and Machine Learning" for large parts of this presentation.

1.1 Preliminaries

Let E and F be $\mathbb{R}$-vector spaces of dimensions n and m respectively. (We want $\dim E = n$ and $\dim F = m$ and not the other way around so that the matrix A which represents f is $m \times n$. Its all very confusing). Let $f : E \to F$ be a linear transformation. Let f be represented by a $m \times n$ matrix A. Let each space be endowed with an inner product, so they are both Euclidean spaces.

For a vector $u \in E$ let $\phi_u : E \to \mathbb{R}$ be the map given by $\phi_u(v) = \langle u, v\rangle$. This is linear by the bilinearity of the inner product

$$\phi_u(v + \lambda w) = \langle u, v + \lambda w\rangle = \langle u, v\rangle + \lambda\langle u, w\rangle = \phi_u(v) + \lambda\phi_u(v)$$

Thus, this is the natural way to associate a vector with a linear functional. Let E^* be the dual space of E.

1. Theorem *The map $\flat : E \to E^*$ given by $u \mapsto \phi_u$ is linear and injective.*

Proof. Linearity follows from bilinearity of $\langle , \rangle$ since for all $w \in E$

$$\flat(u + \lambda v)(w) = \langle u + \lambda v, w\rangle = \langle u, w\rangle + \lambda\langle v, \cdot\rangle = \flat(u)(w) + \lambda\flat(v)(w)$$

which implies $\flat(u + \lambda v) = \flat(u) + \lambda\flat(v)$. Injectivity follows from the positive definiteness of $\langle , \rangle$ since $\flat(u) = 0$ implies $\langle u, v\rangle = 0$ for all v which implies $u = 0$ $\square$

Since E is finite-dimensional, $\flat : E \to E^*$ is a canonical isomorphism. Denote the inverse of bijection $\flat : E \to E^*$ by $\sharp : E^* \to E$.

For a given $v \in F$ and $f : E \to F$ consider the linear functional $\psi_{f,v} : E \to \mathbb{R}$ given by $\psi_{f,v}(u) = \langle f(u), v\rangle$. By theorem 1, there we can represent this linear functional as $\langle u, \sharp(\psi_{f,v})\rangle$ for some element $\sharp(\psi_{f,v}) \in E$. Note that for all $u \in E$

$$\begin{aligned}\langle u, \sharp(\psi_{f,v_1+\lambda v_2})\rangle &= \langle f(u), v_1 + \lambda v_2\rangle = \langle f(u), v_1\rangle + \lambda\langle f(u), v_2\rangle = \langle u, \sharp(\psi_{f,v_1})\rangle + \lambda\langle u, \sharp(\psi_{f,v_1})\rangle \\ &= \langle u, \sharp(\psi_{f,v_1}) + \lambda\sharp(\psi_{f,v_1})\rangle\end{aligned}$$

This shows the mapping $v \mapsto \sharp(\psi_{f,v})$ is linear.

2. Definition The *adjoint* of $f : E \to F$ is the unique linear transformation $f^* : F \to E$ given by $f^*(v) = \sharp(\psi_{f,v})$. It satisfies $\langle f(u), v\rangle = \langle u, f^*(v)\rangle$.

[*]mccorvie@berkeley.edu

Now the mapping $* : \hom(E,F) \to \hom(F,E)$ given by $f \mapsto f^*$ is itself linear since

$$\langle u, (f+\lambda g)^*(v)\rangle = \langle (f+\lambda g)(u), v\rangle = \langle f(u), v\rangle + \lambda\langle g(u), v\rangle = \langle u, f^*(v)\rangle + \lambda\langle u, g^*(v)\rangle = \langle u, (f^*+\lambda g^*)(v)\rangle$$

For all $u \in E$ and $v \in F$ we have

$$\langle f^{**}(u), v\rangle = \langle u, f^*(v)\rangle = \langle f(u), v\rangle$$

which shows that $f^{**} = f$. Finally note that

$$\langle (f\circ g)(u), v\rangle = \langle f(g(u)), v\rangle = \langle g(u), f^*(v)\rangle = \langle u, g^*(f^*(v))\rangle$$

from which it follows that $(f\circ g)^* = g^* \circ f^*$.

3. Corollary *The transformations $f \circ f^*$ and $f^* \circ f$ are self-adjoint*

In matrix terms, we can write $\langle u, v\rangle$ as $u^\top v$ where u and v are both column vectors, so the transpose $u^\top$ is a row vector. Then if A represents the matrix of a linear transformation f, since $(Au)^\top v = u^\top A^\top v = u^\top(A^\top v)$, the matrix which represents f^* is given by the transpose $A^\top$. Thus we'll write A^* for $A^\top$ which is the matrix which represents the adjoint transformation.

4. Theorem (Spectral decomposition) *Let A be a $n\times n$ matrix with entries in $\mathbb{R}$. Then $A = A^\top$ if and only if A can be written $A = U\Lambda U^\top$ where Λ is a diagonal $n\times n$ matrix and U is a orthogonal $n \times n$ matrix (that is $UU^\top = I$).*

For a proof see, for example, [2].

5. Theorem (Courant-Fischer min-max principle) *Let S be a symmetric $n\times n$ matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ then*

$$\begin{aligned}\sigma_k &= \min_{\substack{U\subset\mathbb{R}^n \\ \dim U = n-k+1}} \max_{x\in U} \frac{x^\top S x}{x^\top x} \\ &= \max_{\substack{U\subset\mathbb{R}^n \\ \dim U = k}} \min_{x\in U} \frac{x^\top S x}{x^\top x}\end{aligned}$$

Proof. The ratio $x^\top Sx/x^\top x$ is invariant under the transformation $x \to \alpha x$ for any scalar α, so WLOG we can restrict attention to vectors with $\|x\| = 1$, in which case the ratio is just $x^\top S x$.

Express x in terms of the basis $u_1, \dots, u_n$ of eigenvectors $x = x_1u_1 + \cdots + x_nu_n$. Then

$$\frac{x^\top S x}{x^\top x} =$$

□

1.2 Singular values

In this section we'll fix a basis for E and F and ignore the distinction between a linear transformation and its matrix representatin A.

By corollary 3, linear transformations A^*A and AA^* are self-adjoint. Therefore these matrices can be diagonalized by an orthogonal matrix and they have real eigenvalues.

6. Lemma *The eigenvalues of A^*A and AA^* are nonnegative*

Proof. Suppose u is an eigenvector of A^*A with eigenvalue λ. Then

$$\lambda\langle u, u\rangle = \langle A^*Au, u\rangle = \langle Au, Au\rangle \geq 0$$

Since $\langle u, u\rangle > 0$ the claim follows. $\square$

7. Lemma *If $f : E \to F$ is represented by matrix A and $g : F \to E$ is represented by matrix B then AB and BA have the same non-zero eigenvalues*

Proof. Consider

$$X = \begin{pmatrix} \lambda I_m & A \\ B & I_n \end{pmatrix} \qquad \text{and} \qquad Y = \begin{pmatrix} I_m & 0 \\ -B & \lambda I_n \end{pmatrix}$$

It follows that

$$XY = \begin{pmatrix} \lambda I_m - AB & \lambda A \\ 0 & \lambda I_n \end{pmatrix} \qquad \text{and} \qquad YX = \begin{pmatrix} \lambda I_m & A \\ 0 & \lambda I_n - BA \end{pmatrix}$$

Since these products are upper block triangular matrices, we can read off $\det XY = \lambda^n \det(\lambda I_m - AB)$ and $\det YX = \lambda^m \det(\lambda I_n - BA)$. Therefore these polynomials in λ are equal, which shows the characteristic polynomials of $\chi_{AB}(\lambda)$ and $\chi_{BA}(\lambda)$ are the same, up to factors of λ. So the non-zero eigenvalues of AB and BA are equal, and have equal multiplicity. $\square$

8. Definition The eigenvalues of A^*A can be written $\sigma_1^2, \dots \sigma_n^2$ for non-negative real numbers $\sigma_1, \dots, \sigma_n$. The positive σ_i are called the *singular values* of A. By lemma 7, A^* and A have the same singular values.

9. Proposition (Variational Characterization)

$$\begin{aligned} \sigma_k &= \min_{\substack{S \subset \mathbb{R}^n \\ \dim S = n-k+1}} \max_{\substack{x \in S \\ \|x\|_2 = 1}} \|Ax\|_2 \\ &= \max_{\substack{S \subset \mathbb{R}^n \\ \dim S = k}} \min_{\substack{x \in S \\ \|x\|_2 = 1}} \|Ax\|_2 \end{aligned}$$

Proof. This is a direct corollary of theorem $\square$

10. Proposition *The matrix A and its adjoint A^* have the following properties*

1. $\operatorname{Ker} A = \operatorname{Ker} A^*A$ *and* $\operatorname{Ker} A^* = \operatorname{Ker} AA^*$
2. $\operatorname{Ker} A = (\operatorname{Im} A^*)^\perp$ *and* $\operatorname{Ker} A^* = (\operatorname{Im} A)^\perp$
3. $\dim \operatorname{Im} A = \dim \operatorname{Im} A^*$
4. A, A^*, A^*A, AA^* *all have the same rank.*
5. *Let u_k be the orthornormal basis of eigenvectors of A^*A, where u_k is associated with eigenvalue σ_k^2. Then the vectors Au_i are orthogonal and have length σ_i*

Proof. 1. For the first equality, clearly if $Au = 0$ then $A^*Au = 0$. Conversely if $A^*Au = 0$ then $\langle Au, Au\rangle = \langle A^*Au, u\rangle = 0$ and hence $Au = 0$ since the inner product is positive definite. The second equality follows from the first substituting A^* for A.

2. From the definition of the adjoint $\langle Au, v\rangle = \langle u, A^*v\rangle$ for all $u \in E$ and $v \in F$ so $Au = 0$ iff $Au \perp v$ for all $v \in F$ iff $u \perp A^*v$ for all $v \in F$ iff $u \in (\operatorname{Im} A^*)^\perp$. The second equality follows from the first substituting A^* for A.

3. Note $\dim(\operatorname{Im} A)+\dim(\operatorname{Ker} A) = n$ by the rank-nullity theorem. Also $\dim(\operatorname{Im} A^*)+\dim(\operatorname{Im} A^{*\perp}) = n$ by decomposing E into orthogonal subspaces. By 2 we have $\dim(\operatorname{Ker} A) = \dim(\operatorname{Im} A^*)^\perp$ so we conclude $\dim \operatorname{Im} A = \dim \operatorname{Im} A^*$.

4. The equality $\operatorname{rank} A = \operatorname{rank} A^*$ is just a restatement of 3. Since $\operatorname{rank} A^*A + \dim \operatorname{Ker} A^*A = n = \operatorname{rank} A + \dim \operatorname{Ker} A$ so $\operatorname{rank} A = \operatorname{rank} A^*A$ follows at once from 1. Substituting A^* for A we get $\operatorname{rank} A^* = \operatorname{rank} AA^*$ showing all these ranks are equal.

5. Note $\langle Au_i, Au_j\rangle = \langle A^*Au_i, u_j\rangle = \sigma_i^2\langle u_i, u_j\rangle = \sigma_i^2\delta_{ij}$. So if $i \neq j$ then Au_i and Au_j are orthogonal and if $i = j$ we see $||Au_i|| = \sqrt{\sigma_i^2} = \sigma_i$

$\square$

1.3 The Main Theorems

11. Theorem (Singular Value Decomposition) *Let f have singular values $\sigma_1, \ldots, \sigma_r$. Given $f : E \to F$ we can decompose into orthogonal direct sums $E = E' \oplus \ker f$ and $F = F' \oplus \ker f^*$ where E' has an orthonormal basis $u_1, \ldots, u_r$ and F' has orthonormal basis $v_1, \ldots, v_r$ and*

$$f(u_k) = \sigma_k v_k \qquad \text{and} \qquad f^*(v_k) = \sigma_k u_k \qquad \text{for } k \leq r$$

Letting $u_{r+1}, \ldots, u_n$ be an orthonormal basis for $\ker f$ and $v_{r+1}, \ldots, v_m$ be an orthonormal basis for $\ker f^*$ we have

$$f(u_k) = 0 \qquad \text{and} \qquad f(v_k) = 0 \qquad \text{for } k > r$$

Proof. We refer repeatedly to proposition 10. Let $u_1, \ldots, u_n$ be the orthonormal basis of eigenvectors for $f^* \circ f$ where $u_1, \ldots, u_r$ are associated with the singular values $\sigma_1, \ldots, \sigma_r$ and $u_{r+1}, \ldots, u_n$ are the basis for $\ker f^* \circ f$. By 1, $\ker f^* \circ f = \ker f$ so $f(u_k) = 0$ for $k = r+1, \ldots, n$. For $k = 1, \ldots, r$ let $v_k = \frac{1}{\sigma_k} f(u_k)$. By (4) the v_k are orthonormal and they span $\operatorname{Im} f$ so they form a basis for $\operatorname{Im} f$. By definition they satisfy $f(u_k) = \sigma_k v_k$. Furthermore $f^*(v_k) = \frac{1}{\sigma_k}(f^* \circ f)(u_k) = \frac{\sigma_k^2}{\sigma_k} u_k = \sigma_k u_k$. Now extend the v_k to an orthonormal basis for all of F by adding $v_{r+1}, \ldots, v_m$, an orthonormal basis for $(\operatorname{Im} f)^\perp$. By (2) we have $f^*(v_k) = 0$ for these new vectors with $k = r+1, \ldots, m$. $\square$

We constructed the v_k from the image of the eigenvectors u_k of $f^* \circ f$, but given the decomposition there is a duality. The v_k are the eigenvectors of $f \circ f^*$ and the u_k could be constructed from their image.

12. Theorem (Matrix SVD) *For every real $m \times n$ matrix A, there are two orthogonal matrices U ($n \times n$) and V ($m \times m$) and a diagonal $m \times n$ matrix D such that $A = VDU^\top$, where D is of the form*

$$D = \begin{pmatrix} \sigma_1 & & \cdots & \\ & \sigma_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \sigma_n \\ 0 & & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & \cdots & 0 \end{pmatrix} \qquad \text{or} \qquad D = \begin{pmatrix} \sigma_1 & & \cdots & & 0 & \cdots & 0 \\ & \sigma_2 & \cdots & & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \sigma_m & 0 & \cdots & 0 \end{pmatrix}$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of A and $\sigma_{r+1} = \cdots = \sigma_{\min(m,n)} = 0$. The columns of U are the eigenvectors of $A^\top A$ and the columns of V are the eigenvectors of $AA^\top$.

Proof. Let u_k, v_k, σ_k be as in theorem 11. Perform a change of basis on E to $u_1, \ldots, u_n$ and a change of basis on F to $v_1, \ldots, v_m$. In terms of this new basis, the linear transformation f is represented by the matrix $A' = V^{-1}AU = V^\top AU$ where U has columns u_k and V has columns v_k. (Here $V^{-1} = V^\top$ since V is orthogonal by virtue of the fact the v_k are orthonormal). By theorem 11, f has a simple form in terms of these bases, and given by $A' = D$. From this we immediately conclude $A = VDU^{-1} = VDU^\top$ is a decomposition of A in terms of a diagonal matrix and two orthogonal matrices. □

There's another form which is sometimes useful

13. Theorem (Compact Matrix SVD) *For every real $m \times n$ matrix A with singular values $\sigma_1, \ldots, \sigma_r$, there is a $n \times r$ matrix U_0 with orthonormal columns and a $m \times r$ matrix V_0 with orthonormal columns and a $r \times r$ diagonal matrix*

$$D_0 = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{pmatrix}$$

such that $A = V_0 D_0 U_0^\top$

Proof. Using the notation theorem 11, consider A_0 the matrix which represents the transformation $f^0 : E' \to F'$. If E' has basis $u_1, \ldots, u_r$ and F' has basis $v_1, \ldots, v_r$ then $A_0 = D_0$. Let $\pi : E \to E'$ be the projection onto the subspace. In terms of the basis of E is $u_1, \ldots, u_n$, then the projection has the matrix $P = \begin{pmatrix} I_r & 0 \end{pmatrix}$. Let $\iota : F' \hookrightarrow F$ be the inclusion map. In terms of the basis $v_1, \ldots, v_m$ of F, the inclusion has matrix representation $Q = \begin{pmatrix} I_r \\ 0 \end{pmatrix}$.

So $A = VQD_0PU^\top$. If we let $V_0 = VQ$ and $U_0^\top = PU^\top$ we get the desired decomposition. □

14. Corollary *There are orthonormal vectors $u_1, \ldots, u_r$ and $v_1, \ldots, v_n$ and singular values $\sigma_1, \ldots, \sigma_r$ such that $A = \sum_{i=1}^r \sigma_i v_i u_i^\top$*

Proof. Let $A' = \sum_{i=1}^r \sigma_i v_i u_i^\top$. It suffices to show that $A'u = Au$ for every u in some basis of E. This is straitforward to verify for the basis $u_1, \ldots, u_n$ of eigenvalues of A^*A since $A'u_j = \sum_{i=1}^r \sigma_i v_i u_i^\top u_j = \sum_{i=1}^r \sigma_i \delta_{ij} v_i = \sigma_j v_j$. □

This representation gives rise to the "low-rank approximation" where we truncate this expansion to remove terms with small singular values.

1.4 Eigenvalues and Singular Values

Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 \\ 0 & 1 & 2 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & 2 \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \qquad A^\top A = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 & 0 \\ 2 & 5 & 2 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 5 & 2 \\ 0 & 0 & 0 & \ldots & 2 & 5 \end{pmatrix}$$

($A^\top A$ is tridiagonal). The eigenvalues of A are all 1, but the singular values have a wide spread with $\sigma_1/\sigma_n = \mathrm{cond}_2(A) \geq 2^{n-1}$

Suppose A is a complex square matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ and singular values $\sigma_1, \dots, \sigma_n$. (Note that previously I only considered real matrices, but all the results carry over with the proper modifications). In general, the singular values and the eigenvalues of a complex square matrix can be quite different. However, they are related by

$$\sigma_1^2 \cdots \sigma_n^2 = \det(A^*A) = |\det(A)|^2 = |\lambda_1|^2 \cdots |\lambda_n|^2$$

so we have $|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n$.

Even if A is Hermitian, the singular values and the eigenvalues will be different, if A has negative eigenvalues. For example, $-I = (I) \cdot (-I) \cdot (I)$ is a valid spectral decomposition, but not a valid SVD. On the other hand $-I = (-I) \cdot (I) \cdot I$ is a valid SVD.

15. Proposition (Weyl inequalities) *For any complex $n \times n$ matrix A if $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ are the eigenvalues of A and $\sigma_1, \dots, \sigma_n \in \mathbb{R}^+$ are the singular values of A, listed so that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ then*

$$\begin{aligned} |\lambda_1| \cdots |\lambda_n| &= \sigma_1 \cdots \sigma_n \\ |\lambda_1| \cdots |\lambda_r| &\leq \sigma_1 \cdots \sigma_r \qquad \textit{for } r < n \end{aligned}$$

For a proof see [1]

1.5 Polar Form

16. Definition A pair (R, S) such that $A = RS$ with R orthogonal and S symmetric and positive semidefinite is called a *polar decomposition of A*

17. Proposition *Every matrix $A \in M_{m,n}(\mathbb{R})$ has a unique polar form*

Proof. Let the SVD of A be given by $A = VDU^\top$. Letting $R = VU^\top$ and $S = UDU^\top$ gives the polar form.

TODO uniqueness □

Its straightfoward to go the other way, and express the SVD in terms of the polar form. Given $A = RS$ let $UDU^\top$ be the spectral decomposition of S. Then if we let $V = RU$ we get $A = VDU^\top$ where V and U are orthogonal and D is diagonal and positive seimidefinite.

1.6 Matrix Norms

18. Definition A *matrix norm* is a norm on the space of square $n \times n$ matrices $M_n(\mathbb{R})$ is a norm on the vector space $M_n(K)$ (and hence it satisfies positivity, homogeneity and the triangle inequality) which also satisfies $\|AB\| \leq \|A\|\|B\|$ for all $A, B \in M_n(K)$

19. Definition The *spectral radius* of a square matrix $A \in M_n(\mathbb{R})$ is the magnitude of the maximum eigenvalue $\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$

20. Proposition *For any matrix norm $\|\cdot\|$ on $M_n(\mathbb{R})$ we have $\rho(A) \leq \|A\|$*

Proof. First we assume $\|\cdot\|$ is a norm over $M_n(\mathbb{C})$. Let λ be an eigenvalue of A associated with eigenvector u and let U be the $n \times n$ matrix whose columns are all equal to u. Then since $AU = \lambda U$

$$|\lambda|\|U\| = \|\lambda U\| = \|AU\| \leq \|A\|\|U|$$

so $|\lambda| \leq \|U\|$ for any eigenvalue λ. Maximizing over all eigenvalues gives the result.

In the case of a matrix norm over real matrices, take any matrix norm $\|\cdot\|_c$ over $M_n(\mathbb{C})$ and not it is also a matrix norm over $M_n(\mathbb{R})$. Since all norms of a finite dimensional vector space (like $M_n(\mathbb{R})$)

are equivalent, we have $\|A\|_c \leq C\|A\|$ all $A \in M_n(\mathbb{R})$ for some constant C. Now $\rho(A^k) = \rho(A)^k$ since the eigenvalues of A^k are λ^k where λ is an eigenvalue of A. Therefore

$$\rho(A)^k = \rho(A^k) \leq \|A^k\|_c \leq C\|A^k\| \leq C\|A\|^k$$

Taking kth roots we get $\rho(A) \leq C^{1/k}\|A\|$, and letting $k \to \infty$, we get the desired inequality. ☐

21. Definition The *Frobenius norm* $\|\cdot\|_F$ is the norm $\|\cdot\|_2$ thinking of $M_n(\mathbb{R})$ as the vector space $\mathbb{R}^{n^2}$, which can be written as

$$\|A\|_F = \left(\sum_{i,j=1}^{n} a_{ij}^2\right)^{1/2} = \sqrt{\operatorname{tr}(AA^*)} = \sqrt{\operatorname{tr}(A^*A)}$$

22. Proposition *Given a norm on $\mathbb{R}^n$, there is an induced matrix norm called the* operator norm *$\|\cdot\|_{op}$ given by the function*

$$\|A\|_{op} = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{R}^n, \|x\|=1} \|Ax\|$$

Different norms on $\mathbb{R}^n$ give rise to different matrix norms. The operator norm corresponding to $\|\cdot\|_2$ is called the spectral norm*.*

23. Definition For any matrix $A \in M_{m,n}(\mathbb{C})$ let $\sigma_1 \geq \cdots \geq \sigma_r$ be the singular values. For any $1 \leq k \leq \min(m,n)$ and $p \geq 1$

$$N_{k,p} = (\sigma_1^p + \cdots + \sigma_k^p)^{1/p}$$

(We take $\sigma_k = 0$ for $k > r$ even though its not strictly a singular value since its not positive. It doesn't really matter because it doesn't change the formula). This is called the *Ky Fan p-k-norm*. When $p = 1$ this is called the *Ky Fan k-norm*. When $k = \min(m,n)$ then $N_{k,p}$ is called the *Schatten p-norm*.

When $m = n$ the Ky Fan norms are matrix norms.

24. Proposition *1. The spectral norm $\|\cdot\|_2$ is $\sqrt{\rho(A^*A)} = \sigma_1 = N_1(A)$*

2. The Frobenius norm $\|\cdot\|_F$ is $\sqrt{\operatorname{tr}(A^*A)} = (\sigma_1^2 + \cdots + \sigma_q^2)^{1/2} = N_{q,2}$ (where $q = \min(m,n)$).

3. The trace norm $\operatorname{tr}((A^*A)^{1/2}) = \sigma_1 + \cdots + \sigma_q = N_q(A)$

Proof. 1. Note $\|Au\|_2 = \langle Au, Au\rangle = \langle u, A^*Au\rangle$, so the square of the operator norm is just the Reyleigh quotient for the self-adjoint matrix A^*A. As such it equals the maximum eigenvalue of A^*A, which is $\rho(A^*A) = \sigma_1^2$. Taking square roots yields the desired equation

2. The eigenvalues of A^*A are $\sigma_1^2, \ldots, \sigma_r^2$, so $\sqrt{\operatorname{tr} A^*A} = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}$ as desired

3. In terms of the SVD of $A = VDU^\top$, note that $A^*A = UD^2U^\top$. Therefore we can write $(A^*A)^{1/2} = UDU^\top$, which has eigenvalues $\sigma_1, \ldots, \sigma_r$. So $\operatorname{tr}((A^*A)^{1/2}) = \sigma_1 + \cdots + \sigma_r$ as desired.

☐

For a proof see [1]

1.7 Low-rank approximation

25. Proposition *Let A be an $m \times n$ matrix of rank r and let $VDU^\top = A$ be an SVD for A. Write u_i for the columns of U, v_i for the columns of V, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ for the singular values of*

A. Then a matrix of rank $k < r$ closest to A (in the $\|\cdot\|_2$ norm) is given by

$$A_k = \sum_{i=1}^{k} \sigma_i v_i u^\top = V \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) U^\top$$

and $\|A - A_k\|_2 = \sigma_{k+1}$

Proof. TODO □

What about the Frobenius norm? See Gallier and Quaintance problem 23.4

2 Least Squares Approximation

2.1 The Moore-Penrose Pseudoinverse

26. Definition For a $m \times n$ real matrix A with SVD expansion $A = \sum_{i=1}^{r} \sigma_i v_i u_i^\top$, the *Moore-Pentrose pseudo-inverse* is given by

$$A^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} u_i v_i^\top$$

Equivalently, given SVD $A = VDU^\top$, we have $A^+ = UD^+V^\top$ where D^+ is a $n \times m$ diagonal matrix whose entries are $\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0$.

27. Lemma *$A^+A = \sum_{i=1}^{r} u_k u_k^\top$ and $AA^+ = \sum_{i=1}^{r} v_k v_k^\top$*

Proof. By direct computation

$$A^+A = \sum_{i=1}^{r} \sum_{j=1}^{r} \frac{\sigma_j}{\sigma_i} u_i v_i^\top v_j u_j^\top = \sum_{i=1}^{r} \sum_{j=1}^{r} \frac{\sigma_j}{\sigma_i} \delta_{ij} u_i u_j^\top = \sum_{i=1}^{r} u_k u_k^\top$$

A similar calculation gives the expression for AA^+, □

More explicitly, we have

$$A^+Au_k = \begin{cases} u_k & k \leq r \\ 0 & k > r \end{cases} \qquad AA^+v_k = \begin{cases} v_k & k \leq r \\ 0 & k > r \end{cases}$$

Thus, A^+A is the projection operator onto the orthogonal complement of $\ker A$, which is the same as $\operatorname{Im} A^*$. A similar statement holds for AA^+.

In our notation, $r = \operatorname{rank} A$, and it equals the number of singular values of A. If $r = n$ (which is equivalent to A being injective) then A^+ is a left inverse of A. If $r = m$ (which is equivalent to A being surjective) then A^+ is a right inverse of A. If A is square and nonsingular, then both these conditions hold, and $A^+ = A^{-1}$. In any case, A^+A is the identity on the largest possible subspace of E on which any composition of the form BA could be the identity, namely the orthogonal complement of $\ker A$. Thus A^+ is as "close" to an inverse of A as is possible.

28. Proposition *The Moore-Penrose inverse A^+ satisfies the following properties*

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. *A^+A is symmetric*
4. *AA^+ is symmetric*

Conversely, any matrix which satisfies these properties is the Moore-Penrose inverse.

Proof. To show (1) use the expression for AA^+ in the lemma and perform a similar calculation as the lemma except use the expansion for AA^+ instead of the expansion for A^+. The calculation for (2) is the same, swapping u's for v's.

To show (4), using the expression from the lemma

$$(A^+A)^\top = \left(\sum_{i=1}^r u_k u_k^\top\right)^\top = \sum_{i=1}^r (u_k u_k^\top)^\top = \sum_{i=1}^r u_k u_k^\top = A^+A$$

A similar calculation gives (4).

Now suppose B_1 and B_2 satisfy the four properties above. Note that $BA = (BA)^* = A^*B^*$ where B can be B_1 or B_2. Similarly $AB = B^*A^*$. Therefore

$$B_1 = B_1AB_1 = A^*B_1^*B_1 = A^*B_2^*A^*B_1^*B_1 = B_2AA^*B_1^*B_1 = B_2AB_1AB_1 = B_2AB_1$$

and

$$B_2 = B_2AB_2 = B_2B_2^*A^* = B_2B_2^*A^*B_1^*A^* = B_2AB_2B_1^*A^* = B_2AB_2AB_1 = B_2AB_1$$

Hence $B_1 = B_2$ and the pseudoinverse is unique. □

29. Corollary *If the dimensions of A satisfy $m > \operatorname{rank} A = n$ (so A is tall and has full column rank), then $A^+ = (A^*A)^{-1}A^*$. If the dimensions of A satisfy $n > \operatorname{rank} A = m$ (so A is wide and has full row rank), then $A^+ = A^*(AA^*)^{-1}$*

Proof. This is just a matter of verifying (1)-(4) in proposition 28. First consider the case of a tall matrix A and let $B = (A^*A)^{-1}A^*$. First note $BA = (A^*A)^{-1}A^*A = I$. Clearly this is symmetric and we immediately get $ABA = AI = A$ and $BAB = IB = B$. Finally for $AB = A(A^*A)^{-1}A^*$, using $(M^{-1})^* = (M^*)^{-1}$ its straitforward to verify $(AB)^* = AB$. A similar set of calculations hold for the wide matrix A. □

From property (1) and (2) we see $(A^+A)^2 = A^+A$ and $(AA^+)^2 = AA^+$. Since by (3) and (4) these matrices are symmetric, we see again these are orthogonal projections onto the range of A and A^* respectively.

3 Least Squares regression

Suppose we have an overdetermined set of linear equations $Ax = b$ so that $m > n$. We want to find the closest approximation to a solution, $x^* = \arg\min \|Ax - b\|_2$. In this section we'll always be using this norm, so we'll drop the subscript

Expanding the inner product,

$$\|Ax - b\|^2 = \langle Ax, Ax\rangle - 2\langle Ax, b\rangle + \langle b, b\rangle = x^*A^*Ax - 2x^*A^*b + \|b\|^2$$

This is quadratic in the components of x. Taking gradients, the first-order condition is

$$\nabla_x \|Ax - b\|^2 = 2A^*Ax - 2A^*b = 0$$

or

$$A^*Ax = A^*b$$

These linear equations are called the *normal equation* for least squares.

30. Theorem *Every linear system $Ax = b$ where A is an $m \times n$ real matrix has a unique least squares solution x^+ of smallest norm.*

Proof. Let's consider the geometry of the situation. We can decompose $\mathbb{R}^m = \operatorname{Im} A \oplus (\operatorname{Im} A)^\perp$ into a subspace which is the range of A and its orthogonal complement. I claim there is a unique point in $\operatorname{Im} A$ which is closest to b, and this point is the orthogonal projection of b onto $\operatorname{Im} A$.

Say $b = b^\circ + b^\perp$ where $b^\circ \in \operatorname{Im} A$ and $b^\perp \in (\operatorname{Im} A)^\perp$ is the unique representation of b. Then for any $y \in \operatorname{Im} A$ we have

$$\|y - b\|^2 = \|y - b^\circ - b^\perp\|^2 = \|y - b^\circ\|^2 + \|b^\perp\|^2$$

since $y - b^\circ \in \operatorname{Im} A$ and $b^\perp \in \operatorname{Im} A^\perp$ are orthogonal. Clearly this is minimized when $y = b^\circ$, which occurs when y equals the orthogonal projection of b onto the subspace $\operatorname{Im} A$. Let y^+ denote the least squares element of $\operatorname{Im} A$.

Now it may be there are multiple $x \in \mathbb{R}^n$ such that $y^+ = Ax$. Say x and x' are two solutions. Then note $Ax = y^+ = Ax'$ iff $A(x - x') = 0$ iff $x - x' \in \ker A$, which shows any two solutions are related by a member of $\ker A$. Thus we can decompose the domain of A into orthogonal complements $\mathbb{R}^n = \ker A \oplus \ker A^\perp$, for any solution $Ax = y^+$ we can uniquely write $x = x^\circ + x^\perp$ where $x^\circ \in \ker A$ and $x^\perp \in \ker A^\perp$. Since $\|x\|^2 = \|x^\circ\|^2 + \|x^\perp\|^2$, the unique minimum length vector x^+ satisfying $Ax^+ = y^+$ is the one with no component in $\ker A$. We can find x^+ among all solutions by taking any solution and projecting it onto $\ker A^\perp$. $\square$

The preceding shows that a necessary and sufficient condition for x to minimize $\|Ax - b\|$ is that $Ax - b \in (\operatorname{Im} A)^\perp$. But by proposition 10, this is equivalent to $Ax - b \in \ker A^*$. Therefore, $A^*(Ax - b) = 0$ and we recover the normal equations $A^*Ax = A^*b$.

31. Theorem *The element x^+ described in theorem 30 is given by $x^+ = A^+b$, where A^+ is the Moore-Penrose pseudoinverse of A*

Proof. First assume A is diagonal and we can search for the minimum length solution which minimizes $\|Dx - b\|$. By inspection, $x^+ = (b_1/\sigma_1, \ldots, b_r/\sigma_r, 0, \ldots, 0)^\top$ has the desired properties, since it exactly zeros the first r coordinates in b and the coordinates in positions $r + 1, \ldots, m$ are unaffected by Ax. The minimality of x^+ follows from the fact the x-coordinates in positions $r + 1, \ldots, n$, have no effect so x^+ is minimal when these are all 0. Thus the solution is $x^+ = D^+b$.

In the general case, we can apply an orthogonal transformation to $\mathbb{R}^m$ and $\mathbb{R}^n$ to get an equivalent problem. Let A have SVD $A = VDU^\top$. So, x^+ is the minimum length solution which minimizes $\|Ax - b\|$, if and only if it is also the minimum length solution which minimizes $\|V^\top Ax - V^\top b\|$ for any orthogonal tranformation V. Similarly x^+ is the minimum length solution which minimizes $\|Ax - b\|$ iff $U^\top x^+$ is the minimum length solution which minimizes $\|AUx - b\|$. Here the minimality of $U^\top x$ follows from the fact $\|U^\top x^+\| = \|x^+\|$.

So we find the solution x^+ satisfies $Ux^+ = D^+V^\top b$, and hence $x^+ = UD^+V^\top x = A^+x$ $\square$

For an alternative proof which is less computational, we can verify the normal equations for $x^+ = A^+b$, which are $A^*AA^+b = A^*b$. Let $b = b^\circ + b^\perp$ as above with $b^\circ \in \operatorname{Im} A$ and $b^\perp \in (\operatorname{Im} A)^\perp$. Now AA^+ is the identity on $\operatorname{Im} A$ and 0 on $(\operatorname{Im} A)^\perp$, so $AA^+b = b^\circ$ and $A^*AA^+b = A^*b^\circ$. On the other hand, since $(\operatorname{Im} A)^\perp = \ker A^*$ we have $A^*b = A^*(b^\circ + b^\perp) = A^*b^\circ$, and we've verified the solution. To verify A^+b is minimal length, observe $A^+b \in (\ker A)^\perp = \operatorname{Im} A^*$ since $\operatorname{Im} A^+ = \operatorname{Im} A^*$.

4 Principal Components Analysis

[1] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, first edition, 1994.

[2] Gilbert Strang. *Linear Algebra and its Applications*. Saunders HBJ, third edition, 1988.