线性代数基本定理

目 录

Gilbert Strang (1934 ~) 是 MIT 的数学教授. 从 1999 年开始, 他的线性代数讲课视频陆续发布在 MIT 开放式课程 (OCW) 上, 获得了全球学生和数学爱好者的喜爱, 公开课视频在 Youtube 和 Bilibili 等网站上都有转载, 阅览量超过 1000 万次, 这使他一度成为互联网教学的 "明星". 麻省理工学院开放课程 (OCW) 主任柯特 · 牛顿说:"他一直是开放课程里的宠儿, 这是大家公认的." 他的线性代数教材也被翻译为法语, 德语, 希腊语, 日语和葡萄牙语, 在国内也于 2019 年出版了英文原版第五版.

他在 1993 年发表一篇关于线性代数基本定理的教学论文, 该文以 SVD 为核心, 强调了从四个子空间的角度来直观理解求解线性方程组 $Ax = b$, 内容涉及高斯消元法, 正交互补子空间, 最小二乘法, 相似对角化, 特征值和特征向量, 奇异值分解, 伪逆, 投影等重要基本概念. 编者在学习的过程中顺便把这篇短文做了录入排版, 重新绘制了所有插图, 修改了几处小的印刷错误, 并根据自己的理解对文中相关的知识点做了一些脚注.

— 编者 (5070319@qq.com)

2021.05.21 ~ 2021.05.31

The Fundamental Theorem of Linear Algebra [1]

Gilbert Strang

1 Four subspaces

This paper is about a theroem and the pictures that go with it. The theorem describes the action of an m by n matrix. The matrix A produces a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ — but this picture by itself is too large. The "truth" about $Ax = b$ is expressed in terms of four subspaces (two of $\mathbb{R}^n$ and two of $\mathbb{R}^m$). The pictures aim to illustrate the action of A on those subspaces, in a way that students won't forget.

The First step is to see[2] Ax as a *combination of the columns of* A. Until then the multiplication Ax is just numbers. This step raises the viewpoint to subspaces. We see Ax in the **column space**. Solving $Ax = b$ means finding all combinations of the columns that produce b in the column space:

$$\underset{\text{Columns of } A}{\begin{bmatrix} | & | & \cdots & | \\ | & | & & | \end{bmatrix}} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1(\text{column 1}) + x_2(\text{column 2}) + \cdots + x_n(\text{column n}) = b.$$

The column space is the range $R(A)$, a subspace of $\mathbb{R}^m$. This abstraction, from entries in A or x or b to the picture based on subspace, is absolutely essential. Note how subspaces enter *for a purpose*. We could invent vector spaces and construct bases at random. That misses the purpose. Virtually all algorithms and all applications of linear algebra are understood by moving to subspaces.

The key algorithm is elimination. Multiples of rows are substracted from other rows (and rows are exchanged). There is no change in the ***row space***. This subspace contains all combinations of the rows of A, which are the columns of A^T. The row space of A is the column space $R(A^T)$.

[1] 原文于 1993 年发表在 *The American Methematical Monthly*, Vol. 100, No. 9. (Nov., 1993), pp. 848-855.

[2] 这个视角属于基视角, 即当 A 中各列构成 $\mathbb{R}^m$ 的一组基时, 把求解 $Ax = b$ 这一逆问题视为, 求标准基下的向量 b 在以 A 的各列作为新基时的坐标 x. 一般地, 当 A 的各列不足以构成 $\mathbb{R}^m$ 的一组基 (列亏秩) 时, A 中的各列可以视为原子 (atom) 或库, 这对应到原文所说的视角, 即寻找原子的各种组合方式以合成 b;

另一种视角是变换视角, 即把 A 视为一个变换 τ, 它作用于 x 并产生 b ($A : x \mapsto b$), 求解 $Ax = b$ 这个逆问题就是确定哪些 x 在 τ 的作用下会变成 b. 实际上, 由于 A 已经是一个矩阵, 它并不是变换 τ 本身, 而只是 τ(在 $\mathbb{R}^m$ 和 $\mathbb{R}^n$ 中分别选定了一组基以后) 的化身. 所以 A 的各列是 τ(的对应于选定的两组基下) 的骨架 (skeleton), 而 x 和 b 是各自所对应的基下的坐标. 此时如果把骨架看做原子, 把 x 看做系数, 这似乎还是回到了上一种视角.

上述基视角可以贯彻到对表达式 $A^{-1}b$ 的理解上. 正如 L. N. Trefethen 在他写的教材 *Numerical Linear Algebra* 中强调的, 不要被 A^{-1} 中的求逆符号阻碍了对 $x = A^{-1}b$ 的理解:

> This point cannot be emphasized too much, so we repeat: $A^{-1}b$ is the vector of coefficients of the expansion of b in the basis of columns of A.

如果把 $A^{-1}b$(以及 Matlab 中的反斜线表达方式 $x = A\backslash b$) 视为 $\frac{b}{A}$, 或许更容易唤起这种理解. 当然, 这些都只是表达方式而已, 选择一种适合自己就好.

另外, 求解 $Ax = b$ 还有一种优化视角. 在压缩感知中, A 一般是矮胖矩阵 (故其必然列亏秩), 求解 $Ax = b$ 的问题转化成一个优化问题, 即搜索这样的组合, 它使用尽可能少的列向量组合出 b, 亦即使得解向量 x 尽量稀疏. 除了存储方面的优势以外, 据Strange 教授在公开课上的讲解, 稀疏向量更有利于人们对其成分的解读和理解. 这使得 ℓ_1 范数变得越来越重要, 因为在优化过程使用 ℓ_1 范数更容易得到稀疏的解.

The other subspace of $\mathbb{R}^n$ is the ***nullspace*** $N(A)$. It contains all solutions to $Ax = 0$. Those solutions are not changed by elimination[3], whose purpose is to compute them. A by-product of elimination is to display the dimensions of these subspaces, which is the first part of the theorem.

The *Fundamental Theorem of Linear Algebra* has as many as four parts. Its presentation often stops with Part 1, but the reader is urged to include Part 2. (That is the only part we will prove — it is too valuable to miss. This is also as far as we go in teaching.) The last two parts, at the end of this paper, sharpen the first two. The complete picture shows the action of A on the four subspaces with the right bases. Those bases come from the singular value decomposition.

The Fundamental Theorem begins with

- Part 1. ***The dimensions of the subspaces.***
- Part 2. ***The orthogonality of the subspaces***[4].

The dimensions obey the most important laws of linear algebra:

$$\dim R(A) = \dim R(A^T) \quad \text{and} \quad \dim R(A) + \dim N(A) = n.$$

When the row space has dimension r, the nullspace has dimension $n - r$. Elimination identifies r pivot variables and $n - r$ free variables. Those variables correspond, in the echelon form, to columns with pivots and columns without pivots. They give the dimension count r and $n - r$. Students see this for the echelon matrix and believe it for A.

The *orthogonality* of those spaces is also essential, and very easy. Every x in the nullspace is perpendicular to every row of the matrix, exactly because $Ax = 0$:

$$Ax = \begin{bmatrix} — & \text{row } 1 & — \\ — & \text{row } 2 & — \\ \vdots & \vdots & \vdots \\ — & \text{row } m & — \end{bmatrix} x = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The first zero is the dot product of x with row 1. The last zero is the dot product with row m. One at a time, the rows are perpendicualr to any x in the nullspace. So x is perpendulcar to all combinations of the rows.

The nullspace $N(A)$ ***is orthogonal to the row space*** $R(A^T)$.

What is the fourth subspace? If the matrix A leads to $R(A)$ and $N(A)$, then its transpose must lead to $R(A^T)$ and $N(A^T)$. The fourth subspace is $N(A^T)$, ***the null space of*** A^T. We need it! The theory of linear algebra is bound up in the connections between row spaces and column spaces. If $R(A^T)$ is

[3] 高斯消元法不改变行空间, 故也不改变和它正交的零空间.

[4] 正交依赖于内积概念的引入 (内积是一种双线性形式). 有了内积就可以定义向量之间的夹角, 于是才可以讨论正交. 如果在内积的基础上再定义范数 (并在范数的基础上定义度量 $d(x,y) = ||x - y||$), 则又可以进一步讨论向量的长度, 距离, 邻域等概念. 有了范数/度量就可以做比较, 或者反过来, 正如马毅在一次presentation中所讲的 (大意): 为了做比较所以引入度量和归一化, 卷积网络中的各种归一化就是为了 (在某个统一的基础上) 做比较.

orthogonal to $N(A)$, then — just by transposing — the column space $R(A)$ is orthogonal to the "left nullspace" $N(A^T)$. Look at $A^Ty = 0$:

$$A^Ty = \begin{bmatrix} \text{column 1 of } A \\ \vdots \\ \text{column } n \text{ of } A \end{bmatrix} y = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Since y is orthogonal to each column (producing each zero), y is orthogonal to the whole column space. The point is that A^T is just as good a matrix as A. Nothing is new, except A^T is n by m. Therefore the left nullspace has dimension $m - r$.

A^Ty = 0 means the same as $y^TA = 0^T$. With the vector on the left, y^TA is a combination of the rows of A. Contrast that with $Ax =$ combination of the columns[5].

2 The First Picture: Linear Equations

Figure 1 shows how A takes x into the column space. The nullspace goes to the zero vector. Nothing goes elsewhere in the left nullspace — which is waiting its turn.

With b in the column space, $Ax = b$ can be solved. There is a *particular* solution x, in the row space. The *homogeneous* solutions x, form the nullspace. The general solution is $x_r + x_n$. The particularity of x_r, is that it is orthogonal to every x_n.

May I add a personal note about this figure? Many readers of *Linear Algebra and Its Applications* [4] have seen it as fundamental. It captures so much about $Ax = b$. Some letters suggested other ways to draw the orthogonal subspaces — artistically this is the hardest part[6]. The four subspaces (and very possibly the figure itself) are of course not original. But as a key to the teaching of linear algebra, this illustration is a gold mine.

[5] 这里涉及到对矩阵乘法从不同角度的理解. 正如回字有四种写法一样, 矩阵乘法 $M = AB$ 的也有四种等价的定义:

a). 行列内积: $m_{ij} = a_{i:}^T b_{:j}$, 这是最常见的定义;

b). 按列右乘: $m_{:j} = Ab_{:j}$, M 的各列都是 A 中列向量的组合; 不管多少个矩阵相乘 $M = ABC \cdots Z$, M 的每一列都属于 A 列空间; 特别地, 当 B 为对角矩阵 Λ 时,$A\Lambda$ 即对 A 的各列分别乘以 λ_i; 令 $E = E_1E_2 \cdots E_n$ 为一系列初等变换的组合, 则 AE 就是对 A 的列做初等变换.

c). 按行左乘: $m_{i:} = a_{i:}^T B$, M 的各行都是 B 中行向量的组合; 不管多少个矩阵相乘 $M = ABC \cdots Z$, M 的每一行都属于 Z 的行空间; 同理, ΛA 即对 A 的各行分别乘以 λ_i, 则 EA 就是对 A 的行做初等变换 (高斯消元法).

d). 行列外积: $M = \Sigma a_{:k} b_{k:}^T$, M 为多个单秩矩阵之和; 在矩阵满秩分解, 谱分解, SVD 的奇异分量展开中都可以看到它的应用.

[6] 编者提供另一种画法如下. 把矩阵也放在图中, 且行空间和列空间分别对应于行和列的两侧, 使得左零空间正好在左边; 而且如果矩阵转置的话图形整个做转置即可. 一个不足之处就是映射箭头的方向不再是正常的从左到右的方向:

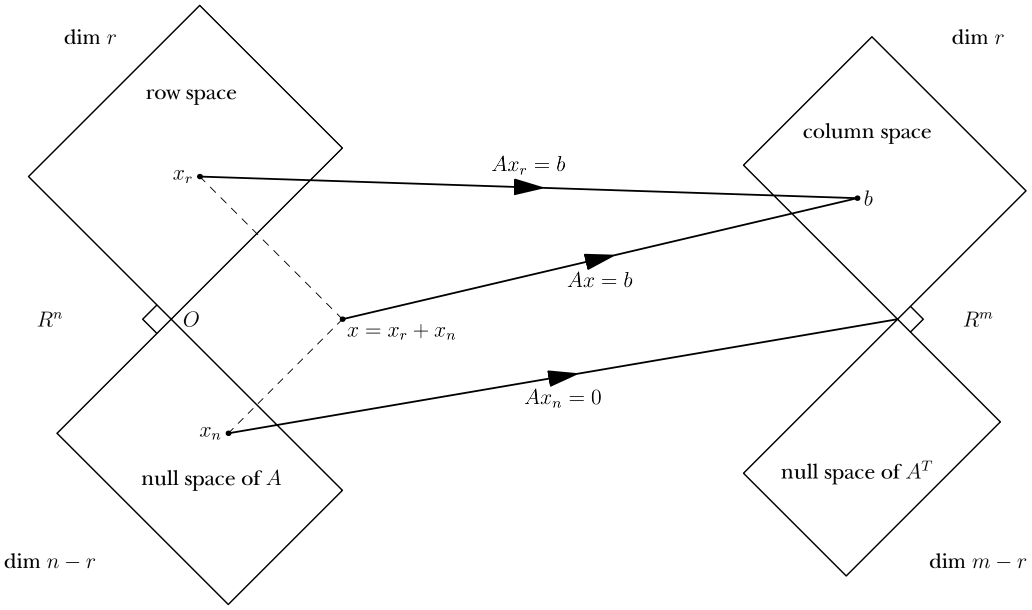
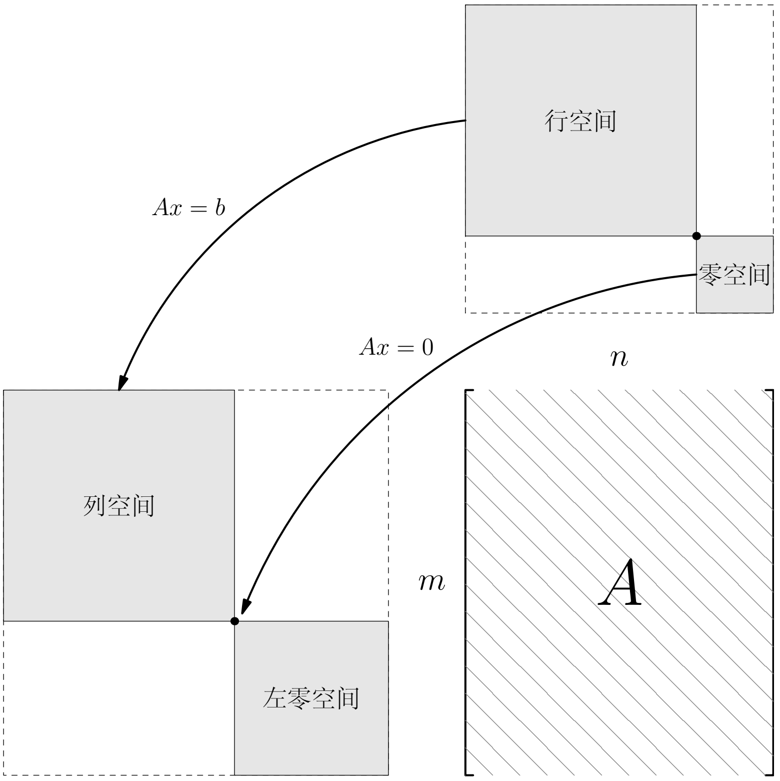


Figure 1. The action of A: Row space to column space, nullspace to zero.

Other writers made a further suggestion. They proposed a lower level textbook, recognizing that the range of students who need linear algebra (and the variety of preparation) is enormous. That new book contains Figures 1 and 2 — also Figure 0[7], to show the dimensions first. The explanation is much more gradual than in this paper—but every course has to study subspaces! We should teach the important ones.



[7] 原文并没有明确标注的 Figure 0，根据文意可能是指前文的公式 dim $R(A)$ = dim $R(A^T)$, dim $R(A)$ + dim $N(A) = n$.

3 The Second Figure: Least Squares Equations

If b is not in the column space, $Ax = b$ cannot be solved. In practice we still have to come up with a "solution". It is extremely common to have more equations than unknowns — more output data than input controls, more measurements than parameters to describe them. The data may lie close to a straight line $b = C + Dt$. A parabola $C + Dt + Et^2$ would come closer. Whether we use polynomials or sines and cosines or exponentials, the problem is still linear in the coefficients C, D, E:

$$\begin{array}{rcl} C + Dt_1 & = & b_1 \\ & \vdots & \\ C + Dt_m & = & b_m \end{array} \qquad \text{or} \qquad \begin{array}{rcl} C + Dt_1 + Et_1^2 & = & b_1 \\ & \vdots & \\ C + Dt_1 + Et_m^2 & = & b_m \end{array} .$$

There are $n = 2$ or $n = 3$ unknowns, and m is larger. There is no $x = (C, D)$ or $x = (C, D, E)$ that satisfies all m equations. $Ax = b$ has a solution only when the points lie exactly on a line or a parabola — then b is in the column space of the m by 2 or m by 3 matrix A.

The solution is to make the error $b - Ax$ as small as possible. Since Ax can never leave the column space, choose the closest point to b in that subspace. This point is the projection p. Then the error vector $e = b - p$ has minimal length.

To repeat: The best combination $p = A\bar{x}$ is the projection of b onto the column space[8]. The error e is perpendicular to that subspace. Therefore $e = b - A\bar{x}$ is in the left nullspace:

$$A^T(b - A\bar{x}) = 0 \quad \text{or} \quad A^TA\bar{x} = A^Tb.$$

Calculus[9] reaches the same linear equations by minimizing the quadratic $\|b - Ax\|^2$. The chain rule just multiplies both sides of $Ax = b$ by A^T.

The "normal equations" are $A^TA\bar{x} = A^Tb$. They illustrate what is almost invariably true — applications that start with a rectangular A end up computing with the square symmetric matrix A^TA[10]. This

[8] 这里指的是正交投影. 一般而言, 一个投影需要定义两个方面: 用 onto 定义向谁投, 用 along 定义顺着谁投. 对于正交投影, 只要定义其中任意一个即可.

[9] 令 a_{ij} 为 A 中第 i 行第 j 列的元素, 有 $S = \|b - Ax\|^2 = \sum_{i=1}^m (b_i - \sum_{j=1}^n a_{ij}x_j)^2$. 分别对 $x_j (j = 1, \cdots, n)$ 求偏导 $\frac{\partial S}{\partial x_j} = 0$, 得到 n 元线性方程组 $A^TAx = A^Tb$, 即为高斯定义的所谓正则方程或法方程 (normal equations).

据陈希孺(1934 ~ 2005) 在《数理统计学简史》第四章中介绍, 最小二乘法的思路在于从整体上平衡超定方程组存在的误差, 和此前多位大数学家 (包括欧拉以及拉普拉斯) 求解超定方程组思路的不同之处在于, 它不使误差过分集中在几个方程内, 而是让它比较均匀地分布于各个方程. 最小二乘法最先由法国数学家勒让德于 1805 年出版, 但是高斯声称自己从 1799 年以来就一直使用该方法. 当时在这两位大数学家之间曾为此发生优先权之争, 其知名度仅次于牛顿和莱布尼兹之间关于微积分发明的优先权之争. 高斯对最小二乘法的最大贡献在于正态误差理论, 以及高斯-马尔科夫定理. 后者从统计学的角度肯定了最小二乘估计的合法性. 在此前, 最小二乘估计只是一种算法, 尽管它看上去合理且有计算简单的优点, 但还不足以回答它在缩小误差这个根本点上, 究竟有何出众之处. 高斯-马尔科夫定理断言, 在线性模型的一切线性无偏估计类中, 最小二乘估计的方差最小. 最小二乘法的缺点在于对异常值的稳定性较差, 对此后人提出对它的修正 (岭估计, 主成分估计) 以及其它替代方法 (最小一乘法) 等.

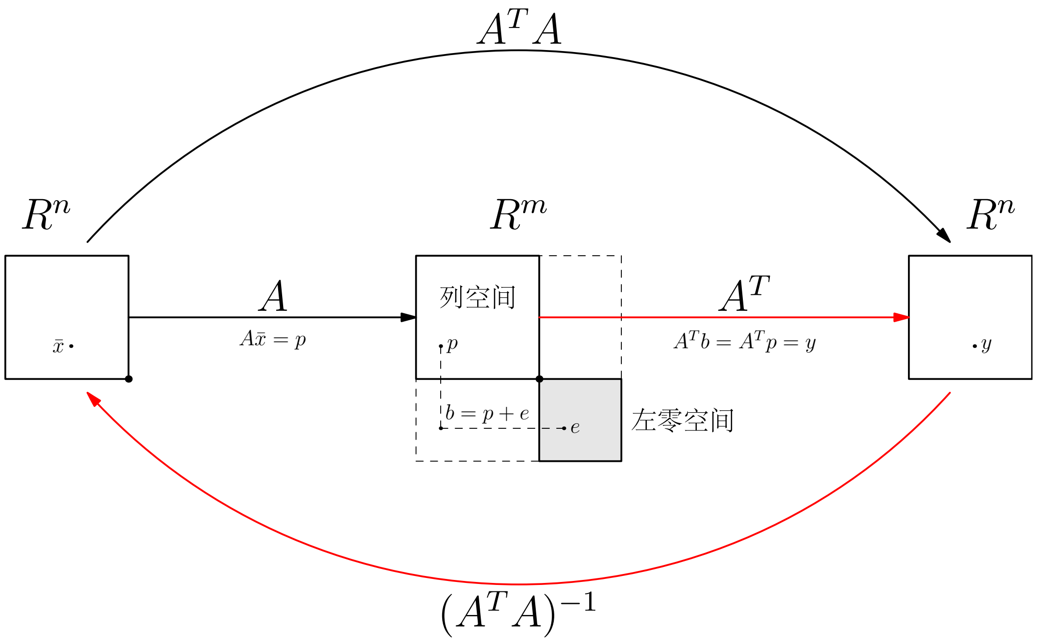
从 $\underset{x}{\operatorname{argmin}}\|b - Ax\|^2$ 可以看出这里默认系数矩阵 A 是确定的, 损失函数是 e(对 b 的修正) 的长度, 这与 e 的几何解释 (投影) 相契合. 另一方面, 解 $\bar{x}$ 固然可以理解成 $A\bar{x} = b - e$, 它也可以理解成 $\bar{A}\bar{x} = b$, 即对 A 做某种修正 ($\bar{A} = A - E$), 使得 $b \in R(\bar{A})$. 后一种理解的问题在于, 对于给定的 $\bar{x}$, 对 A 的修正 E 并不唯一. 考虑到实际应用中系数矩阵 A 和右端项 b 都是观测值 (即都存在随机性误差), 合理的处理好像应该是同时对 A 和 b 做某种修正, 使得 $(A + E)x = b + e$ 成为一个适定方程, 这就是所谓的总体最小二乘法 (Total Least Squares, TLS): $\underset{x,E,e}{\operatorname{argmin}}\|[E : e]\|_F$, 其中 $[E : e]$ 是 E 的增广矩阵, $\|\cdot\|_F$ 为矩阵的 Frobenius 范数.

[10] 对称阵 A^TA 又称为 A(的列向量) 的格拉姆矩阵 (Grammian), 它以丹麦数学家格拉姆 (Jørgen Pedersen Gram, 1850 ~ 1916)命名. 在 $\mathbb{R}$ 上, 格拉姆矩阵是对称阵; 在 $\mathbb{C}$ 上, 格拉姆矩阵是厄米阵 (Grammian is Hermitian). 当且仅当

matrix is invertible provided A has *independent columns*. We make that assumption[11]: The nullspace of A contains only $x = 0$. (Then $A^Tx = 0$ implies $x^TA^TAx = 0$ which implies $Ax = 0$ which forces $x = 0$, so A^TA is invertible.) The picture for least squares shows the action over on the right side — the splitting of b into $p + e$.

A 列满秩时, 其格拉姆矩阵可逆; 格拉姆矩阵半正定; 若 A 列满秩, 则 A^TA 正定.

[11] 对列满秩的瘦矩阵 A, 其零空间坍塌为原点. 从下图中下半部的闭环 $(A^TA)^{-1}A^TA = I_n$ 可以看出红色的部分即 $(A^TA)^{-1}A^T$ 为 A 的左逆, 并且由于 A 列满秩, 使得该左逆具有唯一性, 故进一步赐名为左伪逆 (类似地, 对于行满秩序的胖矩阵可以定义右逆 和右伪逆).



根据上图分析最小二乘法的思路: 对于 $Ax = b$, 由于 $p + e = b \notin R(A)$, 故退而求此次, 假设 b 到 $R(A)$ 的正交投影为 p, 求解 $A\bar{x} = p$. 论文中思路是, 利用左零空间的性质 $A^Te = 0$, 得到 $A^T(b - A\bar{x}) = 0$, 即 $A^TA\bar{x} = A^Tb$. 或者换一个角度看: 利用 A 的左零空间, 用 A^Tb 甩掉了 e, 不过这同时把 p 映射到了 y; 于是只好把左端 Ax 也同时拉下水, 于是得到等式 $A^TA\bar{x} = A^Tb$.

到 $R(A)$ 的正交投影矩阵 P 也可以从上图中直接得到: $P = A(A^TA)^{-1}A^T = AA^+$ (从 $\mathbb{R}^m$ 开始顺着箭头走一圈). 然后根据 $A\bar{x} = Pb$ 也得到同样的结果. 另外, 根据 Strang 教授开公开课上观点, 投影 P 是在尽力向 I 看齐 (I 是一个平凡的投影).

据说高斯当年就用高斯消元法直接求解 $A^TAx = A^Tb$. 从条件数的观点看, 方程 $A^TAx = A^Tb$ 有一个不足之处, 即对比原方程, 条件数增大了: $\kappa(A^TA) = \kappa(A)^2$. 现在一般是用迭代法求 A 的 SVD 分解, 然后得到 $A^+ = V\Sigma^+U^T$, 最后得到 $x^+ = A^+b$.

条件数 $\kappa(A) = \sigma_{\text{max}}/\sigma_{\text{min}}$ 作为矩阵奇异度的指标, 越大则奇异度越高, 方程 (问题) 越不稳定. 方程 $Ax = b$ 的稳定性可以理解成 x 对 A 及 b 的扰动反应的各向均匀性. 以 b 为例, 设对 b 的扰动范围 为一个球, 它对应到 x 的改变范围 就是一个被 Σ^+ 拉伸后的椭球, 这个椭球越椭则说明方程的稳定性越差, 而 $\kappa \geqslant 1$ 就是椭球最长轴和最短轴长度的比值, 用它来刻画椭球的椭度即方程 (问题) 的稳定性 (奇异性).

当 A 列亏秩时, 由于 A^TA 不可逆, 上图中的红色弧线不再存在, 但 $A^TA\bar{x} = A^Tb$ 依然成立. 对照 Figure 5, SVD 把映射缺陷集中放到了 Σ 和 Σ^+: 由 Σ, Σ^+ 充当刽子手 (Σ 负责干掉零空间分量, Σ^+ 负责干掉左零空间分量); 由 U, V 充当三好学生 (双射).

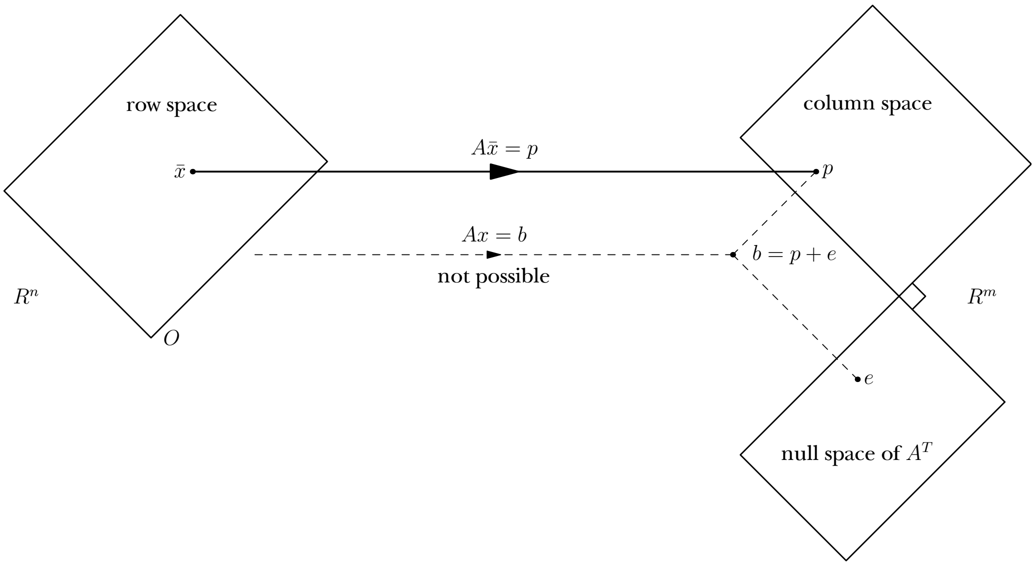


Figure 2. Least squares: $\bar{x}$ minimizes $\|b - Ax\|^2$ by solving $A^TA\bar{x} = A^Tb$.

4 The Third Figure: Orthogonal Bases

Up to this point, nothing was said about *bases for the four subspaces*[12]. Those bases can be constructed from an echelon form — the output from elimination[13]. This construction is simple, but the bases are not perfect. A really good choice, in fact a "canonical choice" that is close to unique, would achieve much more. To complete the Fundamental Theorem, we make two requirements:

- Part 3. ***The basis vectors are orthonormal.***

[12] 在定义了内积的向量空间 V, U 中, 确定了变换 $\tau : V \to U$ 就确定了这里的四个子空间, 它们的存在与 V, U 中基的选择无关. 然而出现了矩阵 A(τ 的化身), 这隐含着在 V, U 中已经各自选择了一组基 $\mathcal{E} = \{e_1, \cdots, e_n\}$ 和 $\mathcal{F} = \{f_1, \cdots, f_m\}$. 只不过它们一般情况下不能正好划分成两部分, 使得一部分是行空间 (列空间) 的基, 另一部分是零空间 (左零空间) 的基. 期望的目标是, 分别对 $\mathcal{E}$ 和 $\mathcal{F}$ 做正交变换, 得到 $\mathbb{R}^n$ 和 $\mathbb{R}^m$ 的另外两组基 $\mathcal{E}'$ 和 $\mathcal{F}'$, 使得它们同时满足下面提到的 Part 3 和 Part 4, 即对 A 做 SVD 分解. 如下图所示, 黑色大写字母 V, U 表示向量空间, 红色希腊字母表示抽象的变换, 蓝色字母表示其对应的矩阵, 矩阵的下标表示它所对应的基; $\mathbb{R}^n$ 和 $\mathbb{R}^m$ 均取标准基; ϕ_1 和 ϕ_2 为 $\mathbb{R}^n$ 到 V 的两个基同构 (分别对应于基 $\mathcal{E}$ 和 $\mathcal{E}'$), ψ_1 和 ψ_2 为 $\mathbb{R}^m$ 到 U 的两个基同构 (分别对应于基 $\mathcal{F}$ 和 $\mathcal{F}'$), ϕ_3 和 ψ_3 为正交变换 (变基); 有 $\tau = \psi_1 \circ \tau_1 \circ \phi_1^{-1} = \psi_2 \circ \tau_2 \circ \phi_2^{-1}$, 且 $\tau_1 = \psi_3 \circ \tau_2 \circ \phi_3^{-1}$ (即 $A = U\Sigma V^T$):

$$\begin{tikzcd}
\mathbb{R}^n \arrow[rrr, dashed, "A_{\mathcal{E},\mathcal{F}}", "\tau_1"'] \arrow[dr, "\phi_1"] & & & \mathbb{R}^m \arrow[dl, "\psi_1"'] \\
& V \arrow[r, dashed, "\tau"] & U & \\
\mathbb{R}^n \arrow[uu, "V_{\mathcal{E}',\mathcal{E}}", "\phi_3"'] \arrow[ur, "\phi_2"'] \arrow[rrr, dashed, "\tau_2", "\Sigma_{\mathcal{E}',\mathcal{F}'}"'] & & & \mathbb{R}^m \arrow[uu, "\psi_3", "U_{\mathcal{F}',\mathcal{F}}"'] \arrow[ul, "\psi_2"]
\end{tikzcd}$$

另一种视角就是直接把 $\mathbb{R}^n$ 视为向量空间, $x \in \mathbb{R}^n$ 为一维数组. 这种情况下, 上图只保留外围的一圈, 左上角和右上角分别采用标准基 $\{e_1, \cdots, e_n\}$ 和 $\{e_1, \cdots, e_m\}$, 左下角和右下角分别采用基 $\{v_1^T, \cdots, v_n^T\}$ 和 $\{u_1^T, \cdots, u_m^T\}$.

SVD 分解的存在性证明是通过构造双线性形式 $z : x, y \mapsto y^tAx (\in \mathbb{R})$, 其中 $(x, y) \in \mathbb{S}^n \times \mathbb{S}^m$ 分别为单位超球面上的点, 由 $\mathbb{S}^n \times \mathbb{S}^m$ 的紧性推出 z 的极大值存在 (可能非唯一). 当 z 取极大值时, 根据内积的性质, y 和 Ax 必定平行, 于是分别取 x 和 y 为 $\mathbb{R}^n$ 和 $\mathbb{R}^m$ 中坐标轴, 然后依此类推可以确定其它所有坐标轴. 当 $r < \max(n, m)$ 时, A 的零空间或左零空间非 0, 此时零空间/左零空间的单位正交基的选择有任意性, 这就是 SVD 变换可能非唯一的原因.

[13] 由于高斯消去法对 A 做行初等变换并不改变行空间, 所以消去法得到的阶梯矩阵 B 中, 非零的行构成了 A 的行空间的一组基; 由于齐次方程组 $Ax = 0$ 和 $Bx = 0$ 有相同的解 (零空间), 而它 (们) 的任意一组解给出了它 (们) 列之间的一种关系, 所以 A 中各列的关系与 B 中各列的关系相同. B 中非主元列必然是该列左侧主元列的线性组合, 即非主元列不独立. 这个关系对 A 同样成立, 所以 A 中由主元所对应的列构成了 A 的列空间的一组基. 此即可以通过消去法得到行空间和列空间的一组基. 零空间 $N(A)$ 的基可以通过解 $Bx = 0$ 得到; 不过左零空间 $N(A^T)$ 的基并不能从 B 中轻易得到, 需要对 A 做增广以后再做消去法, 详见网上讨论.

- Part 4. ***The matrix with respect to these bases is diagonal***[14].

If $v, \cdots, v_r$ is the basis for the row space and $u, \cdots, u_r$ is the basis for the column space, then $Av_i = \sigma_i u_i$. That gives a diagonal matrix Σ. We can further ensure that $\sigma_i > 0$.

Orthonormal bases are no problem — the Gram-Schmidt process is available. But a diagonal form involves eigenvalues. In this case they are the eigenvalues of A^TA and AA^T. Those matrices are symmetric and positive semidefinite, so they have nonnegative eigenvalues and orthonormal eigenvectors (which are the bases!)[15]. Starting from $A^TAv_i = \sigma_i^2 v_i$, here are the key steps:

$$\begin{array}{rlll} v_i^TA^TAv_i = \sigma_i^2 v_i^T v_i & \text{so that} & \|Av_i\| = \sigma_i & \\ AA^TAv_i = \sigma_i^2 Av_i & \text{so that} & u_i = Av_i/\sigma_i & \text{is a unit eigenvector of } AA^T. \end{array}$$

All these matrices have rank r. The r positive eigenvalues σ_i^2 give the diagonal entries σ_i of Σ.

The whole construction is called the *singular value decomposition* (SVD). It amounts to a factorization of the original matrix A into $U\Sigma V^T$, where

1. U is an m by m orthogonal matrix. Its columns $u_1, \cdots, u_r, \cdots, u_m$ are basis vectors for the column space and left nullspace.
2. Σ is an m by n diagonal matrix. Its nonzero entries are $\sigma_1 > 0, \cdots, \sigma_r > 0$[16].
3. V is an n by n orthogonal matrix. Its columns $v_1, \cdots, v_r, \cdots, v_n$ are basis vectors for the row space and nullspace.

The equations $Av_i = \sigma_i u_i$ mean that $AV = U\Sigma$. Then multiplication by V^T gives $A = U\Sigma V^T$.

When A itself is symmetric, its eigenvectors u_i make it diagonal: $A = U\Lambda U^T$. The singular value decomposition extends this spectral theorem to matrices that are not symmetric and not square. The eigenvalues are in Λ, the singular values are in Σ. The factorization $A = U\Sigma V^T$ joins $A = LU$ (elimination) and $A = QR$ (orthogonalization) as a beautifully direct statement of a central theorem in linear algebra.

The history of the SVD is cloudy, beginning with Beltrami [17] and Jordan[18] in the 1870's, but its importance is clear. For a very quick history and proof, and much more about its uses, please see [1]. " The most recurring theme in the book is the practical and theoretical value of this matrix decomposition." The SVD in linear algebra corresponds to the Cartan[19] decomposition in Lie theory [3]. This is one more case, if further convincing is necessary, in which mathematics gets the properties right — and the applications follow.

Example[20]

[14] 即 $Av_i = \sigma_i u_i, i = 1, \cdots, r$, 亦即 $AV = U\Sigma$, 它对应于前面脚注图中的关系 $\tau_1 \circ \phi_3 = \psi_3 \circ \tau_2$.

[15] 假定 $A = U\Sigma V^T$, 易知 $A^TA = V\Sigma^T\Sigma V^T$, $AA^T = U\Sigma\Sigma^TU^T$, 此即对对称阵 $A^TA \in \mathbb{R}^{n\times n}$ 和 $AA^T \in \mathbb{R}^{m\times m}$ 正交相似对角化, 对角阵分别为 $\Sigma^T\Sigma \in \mathbb{R}^{n\times n}$ 和 $\Sigma\Sigma^T \in \mathbb{R}^{m\times m}$, 由单位特征向量系构成的 V 和 U 正好为 $\mathbb{R}^n$ 和 $\mathbb{R}^m$ 的单位正交基. 虽然理论上对 A 的 SVD 分解可由求 A^TA 及 AA^T 的相似对角化完成, 但从条件数的角度考虑, 这并不合算, 实际使用的算法可以直接从 A 得到其 SVD 分解.

[16] 且规定 $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r$.

[17] 意大利数学家贝尔特拉米 (Eugenio Beltrami, 1835 ~ 1900).

[18] 法国数学家约当 (Camille Jordan, 1838 ~ 1922).

[19] 法国数学家埃利·约瑟夫·嘉当 (Élie Joseph Cartan, 1869 ~ 1951), 陈省身的老师.

[20] 原文中有四处数字前面漏掉了负号, 现改.

$$A = \begin{bmatrix}1 & 2\\ 3 & 6\end{bmatrix} = \frac{\begin{bmatrix}-1 & -3\\ -3 & 1\end{bmatrix}}{\sqrt{10}} \begin{bmatrix}\sqrt{50} & 0\\ 0 & 0\end{bmatrix} \frac{\begin{bmatrix}-1 & -2\\ -2 & 1\end{bmatrix}}{\sqrt{5}} = U\Sigma V^T$$

All four subspaces are 1-dimensional. The columns of A are multiples of $\begin{bmatrix}-1\\ -3\end{bmatrix}$ in U. The rows are multiples of $[-1 - 2]$ in V^T .Both A^TA and AA^T have eigenvalues 50 and 0. So the only singular value is $\sigma_1 = \sqrt{50}$.

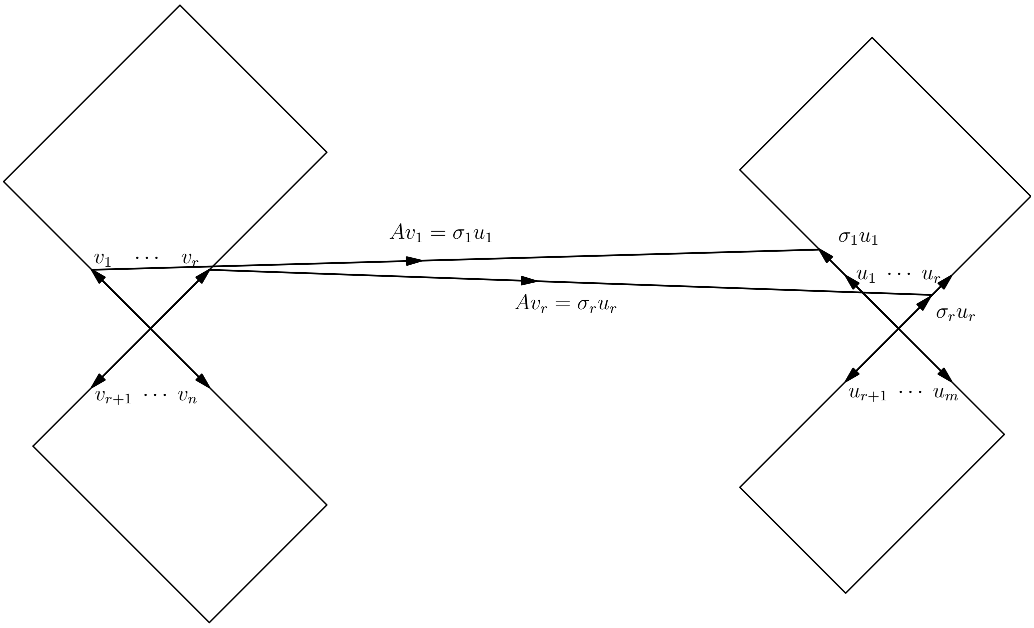


Figure 3. Orthonormal bases that diagonalize A.

The SVD expresses A as a combination of r rank-one matrices:

$$A = U\Sigma V^T = u_1\sigma_1 v_1^T + \cdots + u_r\sigma_r v_r^T \qquad \left(\text{here } A = \begin{bmatrix}1\\ 3\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix}\right).$$

5 The Fourth Figure: The Pseudoinverse

The SVD leads directly to the “pseudoinverse” of A. This is needed, just as the least squares solution $\bar{x}$ was needed, to invert A and solve $Ax = b$ when those steps are strictly speaking impossible. The pseudoinverse A^+ agrees with A^{-1} when A is invertible. The least squares solution of minimum length (having no nullspacc component) is $x^+ = A^+b$. It coincides with $\bar{x}$ when A has full column rank $r = n$ — then A^TA is invertible and Figure 4 becomes Figure 2.

A^+ takes the column space back to the row space [4]. On these spaces of equal dimension r, the matrix A is invertible and A^+ inverts it. On the left nullspace, A^+ is zero[21]. I hope you will feel,

[21] 函数可逆的要求是它能构成双射, 即既是满射又是单射, 其中单射的要求更关键. 一个变换存在零空间就说明它不是单射, 所以在这种意义上, 零空间是令人”讨厌”的东西 (不过在求正交投影时, 零空间却是个合格的”帮凶”, 它会帮你干掉不想要的分量). 正如 Strang 教授在 公开课 Lecture 33: Left and right inverses; pseudoinverse上所讲的:

after looking at Figure 4, that this is the one natural best definition of an inverse. Despite those good adjectives, the SVD and A^+ is too much for an introductory linear algebra course. It belongs in a second course. Still the picture with the four subspaces is absolutely intuitive.

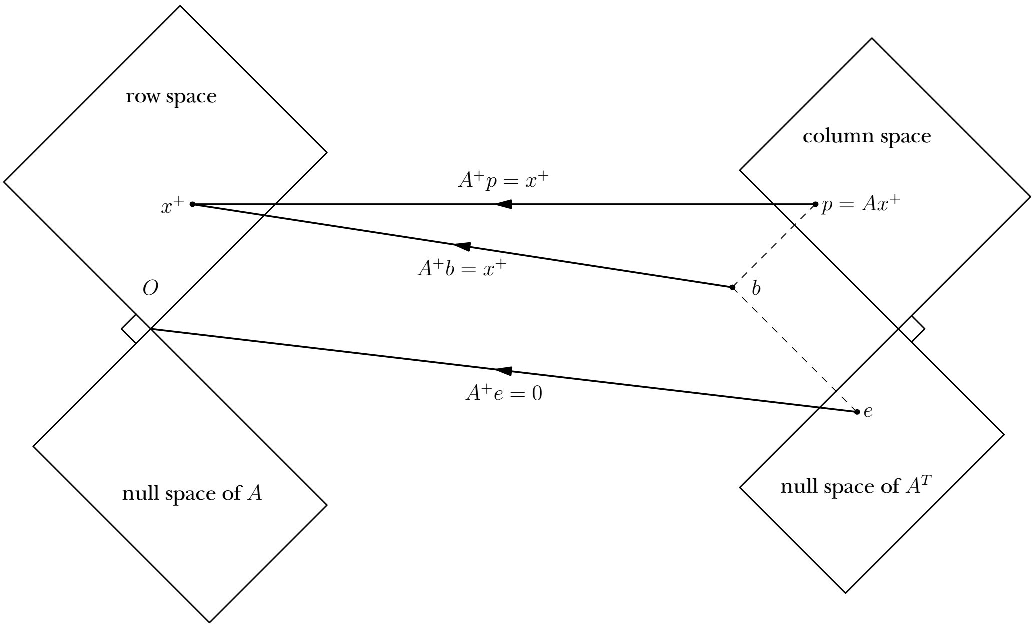


Figure 4. The inverse of A (where possible) is the pseudoinverse A^+.

The SVD gives an easy formula for A^+ , because it chooses the right bases. Since $Av_i = \sigma_i u_i$, the inverse has to be $A^+u_i = v_i/\sigma_i$. Thus the pseudoinverse of Σ contains the reciprocals $1/\sigma_i$. The orthogonal matrices U and V^T are inverted by U^T and V. All together, *the pseudoinverse* of $A = U\Sigma V^T$ is $A^+ = V\Sigma^+U^T$.

Example (continued)

$$A^+ = \frac{\begin{bmatrix} -1 & -2 \\ -2 & 1 \end{bmatrix}}{\sqrt{5}} \begin{bmatrix} 1/\sqrt{50} & 0 \\ 0 & 0 \end{bmatrix} \frac{\begin{bmatrix} -1 & -3 \\ -3 & 1 \end{bmatrix}}{\sqrt{10}} = \frac{1}{50} \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix}$$

Always A^+A is the identity matrix on the row space, and zero on the nullspace:

> If a matrix takes a vector to zero, there is no way its inverse can bring it back to life!

这里对 A 做了一点手术, 即同时限制了它的定义域和值域, 使得它成为一个双射. 故在这个范围内 A 可逆, 它的逆就是 A^+, 而且 A^+ 把左零空间映射到 0. 一般意义上讲, 任意 $f : A \to B$ 都可以被" 标准分解" 成" 满射 + 双射 + 单射", 即: $f = \iota \circ \bar{f} \circ \pi$. 或者说, 这个分解把 f 的" 缺陷" 提取出来分别集中放到了下图的两个地方: $\pi : A \to A/\sim$(非单射缺陷集散地), $\iota : \mathrm{im} f \hookrightarrow B$(非满射缺陷集散地):

$$\begin{tikzcd}
A \arrow[r, two heads, "\pi", "\text{满射}"'] \arrow[rrr, bend left, "f"] & A/\sim \arrow[r, "\bar{f}", "\text{双射}"'] & \mathrm{im} f \arrow[r, hook, "\iota", "\text{单射}"'] & B
\end{tikzcd}$$

如果分别取 $\mathbb{R}^n$ 和 $\mathbb{R}^m$ 对零空间和左零空间的商集, 则可以视 A 在这两商集之间可逆.

$$A^+A = \frac{1}{50}\begin{bmatrix}10 & 20\\ 20 & 40\end{bmatrix} = \text{projection onto the line through } \begin{bmatrix}1\\ 2\end{bmatrix}$$

Similarly AA^+ is the identity on the column space, and zero on the left nullspace:

$$AA^+ = \frac{1}{50}\begin{bmatrix}5 & 15\\ 15 & 45\end{bmatrix} = \text{projection onto the line through } \begin{bmatrix}1\\ 3\end{bmatrix}$$

6 A Summary of the Key Ideas

From its r-dimensional row space to its r-dimensional column space, A yields an invertible linear transformation.

Proof: Suppose x and x' are in the row space, and Ax equals Ax' in the column space. Then $x - x'$ is in both the row space and nullspace. It is perpendicular to itself. Therefore $x = x'$ and the transformation is one-to-one.

The SVD chooses good bases for those subspaces. Compare with the Jordan form for a real square matrix. There we are choosing the *same basis* for both domain and range — our hands are tied. The best we can do is $SAS^{-1} = J$ or $SA = JS$. In general J is not real. If real, then in general it is not diagonal. If diagonal, then in general S is not orthogonal. By choosing *two bases*, not one, every matrix does as well as a symmetric matrix. The bases are orthonormal and A is diagonalized.

Some applications permit two bases and others don't. For powers A^n we need S^{-1} to cancel S . Only a similarity is allowed (one basis). In a differential equation $u' = Au$, we can make one change of variable $u = Sv$. Then $v' = S^{-1}ASv$. But for $Ax = b$, the domain and range are philosophically "not the same space." The row and column spaces are isomorphic, but their bases can be different. And for least squares the SVD is perfect.

This figure[22] by Tom Hern and Cliff Long [2] shows the diagonalization of A. Basis vectors go to basis vectors (principal axes). A circle goes to an ellipse. The matrix is factored into $U\Sigma V^T$. Behind the scenes are two symmetric matrices A^TA and AA^T. So we reach two orthogonal matrices U and V.

[22] 原文中该图没有编号, 现加; 原图的标注意味着 $\sigma_1 < \sigma_2$, 现将二者做了交换, 使得 $\sigma_1 > \sigma_2$; 另外, 红色部分原图没有, 为编者所加.
从伪逆 A^+ 可以马上得到两个正交投影 $P = AA^+$ 和 $P' = A^+A$, 前者是到 A 的列空间的正交投影, 后者是到 A^T 的列空间的正交投影. 特别地, 当 A 为向量 v 时 (即矩阵 A 只有一列, $n = 1$), 其 SVD 分解为 $A = \begin{bmatrix}\frac{v_1}{\rho} & \cdots\\ \vdots & \vdots\\ \frac{v_m}{\rho} & \cdots\end{bmatrix}\begin{bmatrix}\rho\\ 0\\ \vdots\end{bmatrix}1 = v$. 于是 $A^+ = 1\begin{bmatrix}\frac{1}{\rho} & 0 & \cdots\end{bmatrix}\begin{bmatrix}\frac{v_1}{\rho} & \cdots & \frac{v_m}{\rho}\\ \vdots & \vdots & \vdots\end{bmatrix} = \frac{1}{\rho^2}v^T$, 故 $P = AA^+ = \frac{1}{\rho^2}vv^T$, 其中 ρ 为 v 的长度 $\|v\|$. 当 v 为单位向量时有 $P = vv^T$; 同理可得, 当 $A = V$ 且 V 的各列为单位正交向量时, $P = VV^T$. 另外, 由正交投影的性质可知, 到左零空间 (列空间的正交互补空间) 的正交投影 $Q = I - P = I - VV^T = \prod(I - v_iv_i^T)$, 其中 v_i 为 V 的各列. 其几何解释为, 投影 Q 即为了抛弃位于 V 列空间的部分, 它可以通过依次抛弃 V 的列空间的各个一维子空间的分量来达到, 且和顺序无关.

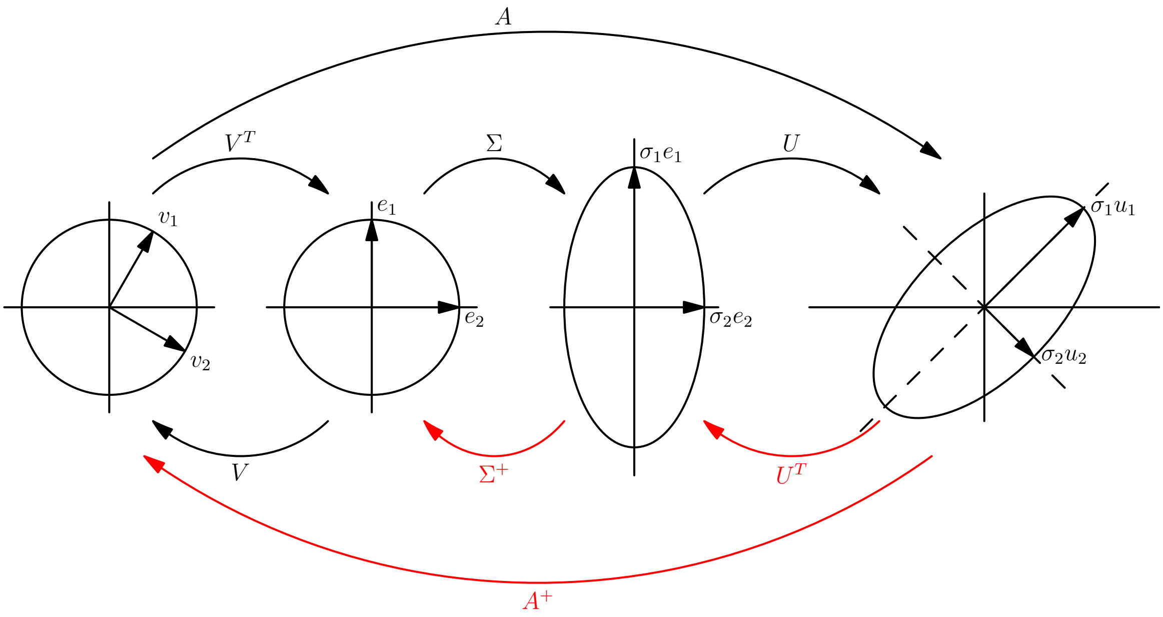


Figure 5. Diagonization of A by SVD.

We close by summarizing the action of A and A^T and A^+:

$$Av_i = \sigma_i u_i \qquad A^T u_i = \sigma_i v_i \qquad A^+ u_i = v_i/\sigma_i \qquad 1 \leqslant i \leqslant r.$$

The nullspaces go to zero. Linearity does the rest.

The support of the National Science Foundation (DMS 90-06220) is gratefully acknowledged.

References

[1] Gene Golub and Charles Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press (1989).

[2] Thomas Hern and Cliff Long, Viewing some concepts and applications in linear algebra, *Visualization in Teaching and Learning Mathematics*, MAA Notes 19 (1991) 173-190.

[3] Roger Howe, Very basic Lie theory, *American Mathematical Monthly*, 90 (1983) 600-623.

[4] Gilbert Strang, *Linear Algebra and Its Applications*, 3rd ed., Harcourt Brace Jovanovich (1988).

[5] Gilbert Strang, *Introduction to Linear Algebra*, Wellesley-Cambridge Press (1993).

Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
gs@math.mit.edu