MATH 680 Computation Intensive Statistics

December 3, 2018

Matrix Basics

Contents

1 Rank (linear algebra)

In linear algebra, the **rank** of a matrix A is the dimension of the vector space generated (or spanned) by its columns. This corresponds to the maximal number of linearly independent columns of A. This, in turn, is identical to the dimension of the space spanned by its rows.

The **column rank** of A is the dimension of the **column space** of A, while the **row rank** of A is the dimension of the **row space** of A.

A fundamental result in linear algebra is that the column rank and the row rank are always equal.

1.1 Proofs that column rank = row rank

Let A be an $m \times n$ matrix with entries in the real numbers whose row rank is r. Therefore, the dimension of the row space of A is r. Let $x_1, x_2, \ldots, x_r$ be a basis of the row space of A. We claim that the vectors $Ax_1, Ax_2, \ldots, Ax_r$ are linearly independent. To see why, consider a linear homogeneous relation involving these vectors with scalar coefficients $c_1, c_2, \ldots, c_r$:

$$0 = c_1 Ax_1 + c_2 Ax_2 + \cdots + c_r Ax_r = A(c_1 x_1 + c_2 x_2 + \cdots + c_r x_r) = Av,$$

where

$$v = c_1x_1 + c_2x_2 + \cdots + c_rx_r.$$

We make two observations:

1. v is a linear combination of vectors in the row space of A, which implies that v belongs to the row space of A,

2. and since $Av = 0$, the vector v is orthogonal to every row vector of A and, hence, is orthogonal to every vector in the row space of A.

The facts 1. and 2. together imply that v is orthogonal to itself, which proves that $v = 0$ or, by the definition of v,

$$c_1x_1 + c_2x_2 + \cdots + c_rx_r = 0.$$

But recall that the x_i were chosen as a basis of the row space of A and so are linearly independent. This implies that

$$c_1 = c_2 = \cdots = c_r = 0.$$

It follows that $Ax_1, Ax_2, \ldots, Ax_r$ are linearly independent.

Now, each Ax_i is obviously a vector in the column space of A. So, $Ax_1, Ax_2, \ldots, Ax_r$ is a set of r linearly independent vectors in the column space of A and, hence, the dimension of the column space of A (i.e., the column rank of A) must be at least as big as r. This proves that row rank of A is no larger than the column rank of A. Now apply this result to the transpose of A to get the reverse inequality and conclude as in the previous proof.

1.2 Properties

We assume that A is an $m \times n$ matrix, and we define the linear map f by

$$f(x) = Ax$$

as above.

- The rank of an $m \times n$ matrix is a nonnegative integer and cannot be greater than either m or n. That is,

  $$\operatorname{rank}(A) \leq \min(m, n).$$

  A matrix that has rank $\min(m, n)$ is said to have full rank; otherwise, the matrix is rank deficient.

- Only a zero matrix has rank zero.

- If B is any $n \times k$ matrix, then

  $$\operatorname{rank}(AB) \leq \min(\operatorname{rank}(A), \operatorname{rank}(B)).$$

- If B is an $n \times k$ matrix of rank n, then

  $$\operatorname{rank}(AB) = \operatorname{rank}(A).$$

- If C is an $l \times m$ matrix of rank m, then

$$\operatorname{rank}(CA) = \operatorname{rank}(A).$$

- **Sylvester's rank inequality**: if A is an m × n matrix and B is n × k, then

$$\operatorname{rank}(A) + \operatorname{rank}(B) - n \leq \operatorname{rank}(AB).$$

- **Subadditivity**:

$$\operatorname{rank}(A + B) \leq \operatorname{rank}(A) + \operatorname{rank}(B)$$

2 Kernel (linear algebra)

The **kernel** (also known as **null space**) of a linear map

$$L : V \to W$$

between two vector spaces V and W, is the set of all elements $\mathbf{v}$ of V for which

$$L(\mathbf{v}) = \mathbf{0},$$

where 0 denotes the zero vector in W. That is, in set-builder notation,

$$\ker(L) = \{\mathbf{v} \in V \mid L(\mathbf{v}) = \mathbf{0}\}.$$

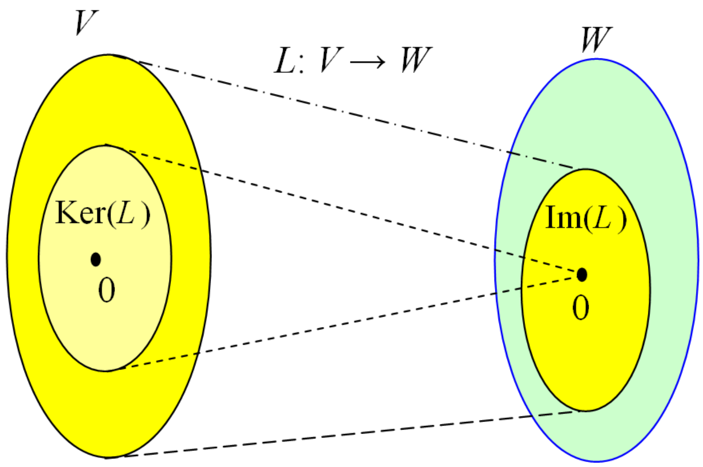

Figure 1: Kernel and image of a map L.

2.1 Properties

The kernel of L is a **linear subspace** of the **domain** V. In the linear map $L : V \to W$, two elements of V have the same **image** in W if and only if their difference lies in the kernel of L:

$$L(\mathbf{v}_1) = L(\mathbf{v}_2) \;\Leftrightarrow\; L(\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}.$$

Consider a linear map represented as a $m \times n$ matrix

$$A \in \mathbb{R}^{m\times n} : \; \mathbb{R}^n \to \mathbb{R}^m$$

with coefficients and operating on column vectors x with n components. The kernel of this linear map is

$$\mathrm{Null}(A) = \ker(A) = \{\mathbf{x} \in K^n | A\mathbf{x} = \mathbf{0}\}.$$

Subspace properties

The kernel of an $m \times n$ matrix A, i.e. the set Null(A), has the following three properties:

1. Null(A) always contains the zero vector, since $A\mathbf{0} = \mathbf{0}$.

2. If $x \in$ Null(A) and $y \in$ Null(A), then $x + y \in$ Null(A). This follows from the distributivity of matrix multiplication over addition.

3. If $x \in$ Null(A) and c is a scalar, then $c\mathbf{x} \in$ Null(A), since $A(c\mathbf{x}) = c(A\mathbf{x}) = c\mathbf{0} = \mathbf{0}$.

The row space of a matrix The product $A\mathbf{x}$ can be written in terms of the dot product of vectors as follows:

$$A^T\mathbf{x} = \begin{bmatrix} \mathbf{a}_1 \cdot \mathbf{x} \\ \mathbf{a}_2 \cdot \mathbf{x} \\ \vdots \\ \mathbf{a}_m \cdot \mathbf{x} \end{bmatrix}.$$

Here $\mathbf{a}_1, \ldots, \mathbf{a}_m$ denote the rows of the matrix A.

It follows that $\mathbf{x}$ is in the kernel of A if and only if $\mathbf{x}$ is orthogonal to each of the row vectors of A. The row space of a matrix A is the span of the row vectors of A. By the above reasoning, the kernel of A is the orthogonal complement to the row space. That is, a vector $\mathbf{x}$ lies in the kernel of A if and only if it is orthogonal to every vector in the row space of A.

Rank–nullity theorem The **rank-nullity theorem** is a fundamental theorem in linear algebra which relates the dimensions of a linear map's **kernel** and **image** with the dimension of its **domain**.

The **kernel** (also known as **null space**) of a linear map

$$A : V \to W.$$

Then

$$\mathrm{Rank}(T) + \mathrm{Nullity}(T) = \dim V$$

where

$$\mathrm{Rank}(T) := \dim \mathrm{Image}\, T$$

$$\mathrm{Nullity}(T) := \dim \mathrm{Kerel}\, T$$

We have

$$A \in \mathbb{R}^{m\times n} : \ \mathbb{R}^m \to \mathbb{R}^n.$$

In the case of an $m \times n$ matrix, the dimension of the domain is n, the number of columns in the matrix. The dimension of the row space of A is called the rank of A, and the dimension of the kernel of A is called the nullity of A. These quantities are related by the rank–nullity theorem

$$\mathrm{rank}(A) + \mathrm{nullity}(A) = n.$$

Nonhomogeneous systems of linear equations The kernel also plays a role in the solution to a nonhomogeneous system of linear equations:

$$A\mathbf{x} = \mathbf{b} \qquad \text{or} \qquad \begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

If $\mathbf{u}$ and $\mathbf{v}$ are two possible solutions to the above equation, then

$$A(\mathbf{u} - \mathbf{v}) = A\mathbf{u} - A\mathbf{v} = \mathbf{b} - \mathbf{b} = \mathbf{0}$$

Thus, the difference of any two solutions to the equation $A\mathbf{x} = \mathbf{b}$ lies in the kernel of A.

It follows that any solution to the equation $A\mathbf{x} = \mathbf{b}$ can be expressed as the sum of a fixed solution $\mathbf{v}$ and an arbitrary element of the kernel. That is, the solution set to the equation Ax = b is

$$\{\mathbf{v} + \mathbf{x} \mid A\mathbf{v} = \mathbf{b},\ \mathbf{x} \in \mathrm{Null}(A)\}\,.$$

2.2 Illustration

We give here a simple illustration of computing the kernel of a matrix. We also touch on the row space and its relation to the kernel.

Consider the matrix

$$A = \begin{bmatrix} 2 & 3 & 5 \\ -4 & 2 & 3 \end{bmatrix}.$$

The kernel of this matrix consists of all vectors $(x, y, z) \in \mathbb{R}^3$ for which

$$\begin{bmatrix} 2 & 3 & 5 \\ -4 & 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which can be expressed as a homogeneous system of linear equations involving x, y, and z:

$$2x + 3y + 5z = 0,$$

$$-4x + 2y + 3z = 0,$$

which can be written in matrix form as:

$$\left[\begin{array}{ccc|c} 2 & 3 & 5 & 0 \\ -4 & 2 & 3 & 0 \end{array}\right].$$

Gauss–Jordan elimination reduces this to:

$$\left[\begin{array}{ccc|c} 1 & 0 & 1/16 & 0 \\ 0 & 1 & 13/8 & 0 \end{array}\right].$$

Rewriting yields:

$$x = -\frac{1}{16}z$$
$$y = -\frac{13}{8}z.$$

Now we can express an element of the kernel:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = c \begin{bmatrix} -1/16 \\ -13/8 \\ 1 \end{bmatrix}.$$

for c a scalar.

Since c is a free variable, this can be expressed equally well as,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = c \begin{bmatrix} -1 \\ -26 \\ 16 \end{bmatrix}.$$

The kernel of A is precisely the solution set to these equations (in this case, a line through the origin in $\mathbb{R}^3$); the vector $(-1, -26, 16)^T$ constitutes a basis of the kernel of A. Thus, the nullity of A is 1.

Note also that the following dot products are zero:

$$\begin{bmatrix} 2 & 3 & 5 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ -26 \\ 16 \end{bmatrix} = 0 \quad \text{and} \quad \begin{bmatrix} -4 & 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ -26 \\ 16 \end{bmatrix} = 0,$$

which illustrates that vectors in the kernel of A are orthogonal to each of the row vectors of A.

These two (linearly independent) row vectors span the row space of A, a plane orthogonal to the vector $(-1, -26, 16)^T$.

With $\mathrm{rank} = 2$, and $\mathrm{nullity} = 1$, and the dimension of A is 3, we have an illustration of the rank-nullity theorem.

3 Matrix Norm

There are three types of matrix norms which will be discussed below:

- Matrix norms induced by vector norms
- Entrywise matrix norms
- Schatten norms.

3.1 Matrix norms induced by vector norms

Suppose a vector norm $\|\cdot\|$ on $\mathbb{R}^d$ is given. Any $m \times n$ matrix A induces a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$ with respect to the standard basis, and one defines the corresponding induced norm or operator norm on the space $\mathbb{R}^{m\times n}$ of all $m \times n$ matrices as follows:

$$
\begin{aligned}
\|A\| &= \sup\{\|Ax\| : x \in \mathbb{R}^n \text{ with } \|x\| = 1\} \\
&= \sup\left\{\frac{\|Ax\|}{\|x\|} : x \in \mathbb{R}^n \text{ with } x \neq 0\right\}.
\end{aligned}
$$

In particular, if the p-norm for vectors $(1 \leq q \leq \infty)$ is used for both spaces $\mathbb{R}^n$ and $\mathbb{R}^m$, then the corresponding induced operator norm is:

$$
\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}.
$$

These induced norms are different from the "entrywise" p-norms and the Schatten p-norms for matrices treated below, which are also usually denoted by $\|A\|_p$.

Any induced operator norm is a sub-multiplicative matrix norm:

$$\|AB\| \leq \|A\|\|B\|$$

this follows from

$$\|ABx\| \leq \|A\|\|Bx\| \leq \|A\|\|B\|\|x\|$$

and

$$\max_{\|x\|=1} \|ABx\| = \|AB\|.$$

Moreover, any induced norm satisfies the inequality

$$\|A^r\|^{1/r} \geq \rho(A), \tag{1}$$

where $\rho(A)$ is the spectral radius of A. For symmetric A, we have equality in 1 for the 2-norm, since in this case the 2-norm is precisely the spectral radius of A.

3.1.1 Special cases

In the special cases of $p = 1, 2, \infty$ the induced matrix norms can be computed or estimated by

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|,$$

which is simply the maximum absolute column sum of the matrix;

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|,$$

which is simply the maximum absolute row sum of the matrix;

$$\|A\|_2 = \sigma_{\max}(A) \leq \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \|A\|_F,$$

where in left hand side $\sigma_{\max}(A)$ represents the largest singular value of matrix A, and on the right hand side $\|A\|_F$ is the Frobenius norm. The first inequality can be derived from the fact that the trace of a matrix is equal to the sum of its eigenvalues. The equality holds if and only if the matrix A is a rank-one matrix or a zero matrix.

In the special case of $p = 2$ (the Euclidean norm or ℓ_2 -norm for vectors), the induced matrix norm is the spectral norm. The spectral norm of a matrix A is the largest singular value of A i.e. the square root of the largest eigenvalue of the positive-semidefinite matrix $A^T A$:

$$\|A\|_2 = \sqrt{\lambda_{\max}\left(A^T A\right)} = \sigma_{\max}(A)$$

where A^T denotes the transpose of A.

3.2 Entrywise matrix norms

These norms treat an $m \times n$ matrix as a vector of size mn , and use one of the familiar vector norms.

For example, using the p-norm for vectors, $p \geq 1$, we get:

$$\|A\|_p = \|\mathrm{vec}(A)\|_p = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^p\right)^{1/p}$$

This is a different norm from the induced p-norm (see above) and the Schatten p-norm (see below), but the notation is the same.

The special case $p = 2$ is the Frobenius norm, and $p = \infty$ yields the maximum norm.

3.2.1 Frobenius norm

When $p = 2$, it is called the Frobenius norm. This norm can be defined in various ways:

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\mathrm{Tr}\left(A^T A\right)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)},$$

where $\sigma_i(A)$ are the singular values of A. Recall that the trace function returns the sum of diagonal entries of a square matrix.

The Frobenius norm has the useful property of being invariant under rotations, that is,

$$\|A\|_{\mathrm{F}}^2 = \|AR\|_{\mathrm{F}}^2 = \|RA\|_{\mathrm{F}}^2$$

for any rotation matrix R. This property follows from the trace definition restricted to real matrices:

$$\|AR\|_{\mathrm{F}}^2 = \mathrm{Tr}\left(R^\top A^\top AR\right) = \mathrm{Tr}\left(RR^\top A^\top A\right) = \mathrm{Tr}\left(A^\top A\right) = \|A\|_{\mathrm{F}}^2$$

and

$$\|RA\|_{\mathrm{F}}^2 = \mathrm{Tr}\left(A^\mathsf{T} R^\mathsf{T} R A\right) = \mathrm{Tr}\left(A^\mathsf{T} A\right) = \|A\|_{\mathrm{F}}^2,$$

where we have used the orthogonal nature of R (that is, $R^\mathsf{T} R = RR^\mathsf{T} = \mathbf{I}$) and the cyclic nature of the trace ($\mathrm{Tr}(XYZ) = \mathrm{Tr}(ZXY)$).

It also satisfies

$$\|A^\mathrm{T} A\|_\mathrm{F} = \|AA^\mathrm{T}\|_\mathrm{F} \leq \|A\|_\mathrm{F}^2$$

and

$$\|A+B\|_\mathrm{F}^2 = \|A\|_\mathrm{F}^2 + \|B\|_\mathrm{F}^2 + 2\langle A, B\rangle_\mathrm{F},$$

where $\langle A, B\rangle_\mathrm{F}$ is the Frobenius inner product.

3.2.2 Frobenius inner product

the Frobenius inner product is a binary operation that takes two matrices and returns a number. It is often denoted $\langle \mathbf{A}, \mathbf{B}\rangle_\mathrm{F}$ $\langle \mathbf{A}, \mathbf{B}\rangle_\mathrm{F}$. The operation is a component-wise inner product of two matrices as though they are vectors. The two matrices must have the same dimension—same number of rows and columns—but are not restricted to be square matrices.

Given two complex number-valued n×m matrices A and B, written explicitly as

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nm} \end{pmatrix}$$

the Frobenius inner product is defined by the following summation Σ of matrix elements,

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}} = \sum_{i,j} A_{ij} B_{ij} \,, = \mathrm{tr}\left(\mathbf{A}^T \mathbf{B}\right)$$

The calculation is very similar to the dot product, which in turn is an example of an inner product.

Properties

$$\langle a\mathbf{A}, b\mathbf{B} \rangle_{\mathrm{F}} = ab\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}}$$

$$\langle \mathbf{A} + \mathbf{C}, \mathbf{B} + \mathbf{D} \rangle_{\mathrm{F}} = \langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}} + \langle \mathbf{A}, \mathbf{D} \rangle_{\mathrm{F}} + \langle \mathbf{C}, \mathbf{B} \rangle_{\mathrm{F}} + \langle \mathbf{C}, \mathbf{D} \rangle_{\mathrm{F}}$$

$$\langle \mathbf{B}, \mathbf{A} \rangle_{\mathrm{F}} = \langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}}$$

For the same matrix,

$$\langle \mathbf{A}, \mathbf{A} \rangle_{\mathrm{F}} \geq 0 \,.$$

The inner product induces the Frobenius norm

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle_{\mathrm{F}}} \,.$$

3.3 Schatten norms

The Schatten p-norms arise when applying the p-norm to the vector of singular values of a matrix. If the singular values are denoted by σ_i, then the Schatten p-norm is defined by

$$\|A\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(A) \right)^{1/p}.$$

These norms again share the notation with the induced and entrywise p-norms, but they are different.

All Schatten norms are sub-multiplicative. They are also unitarily invariant, which means that $\|A\| = \|UAV\|$ for all matrices A and all unitary matrices U and V.

The most familiar cases are $p = 1, 2, \infty$.

- The case $p = 2$ yields the Frobenius norm, introduced before.
- The case $p = \infty$ yields the spectral norm, which is the operator norm induced by the vector 2-norm (see above).
- Finally, $p = 1$ yields the nuclear norm (also known as the trace norm), defined as

$$\|A\|_{\text{Tr}} = \text{Tr}\left(\sqrt{A^T A}\right) = \sum_{i=1}^{\min\{m,n\}} \sigma_i(A).$$

What are the singular values of an $n \times n$ square orthogonal matrix? SVD of a matrix A is $A = U\Sigma V^T$, where U and V are orthogonal and Σ is nonnegative real diagonal.

Now, let X be orthogonal. Note that $X = U\Sigma V^T$, where $U := X$ is orthogonal, $\Sigma :=$ I is diagonal, and $V :=$ I is orthogonal. So, singular values are all equal to 1.

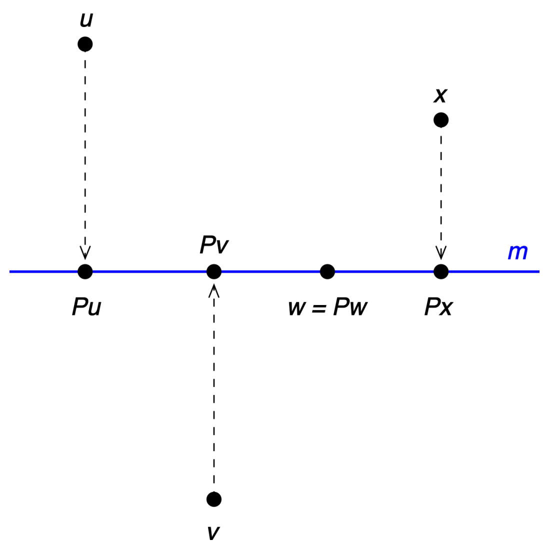

Figure 2: The transformation P is the **orthogonal** projection onto the line m.

How do we know that the set of all orthogonal matrices is nonconvex? If Q is orthogonal, then $-Q$ is also orthogonal, but zero matrix is not orthogonal.

4 Projection

A **projection** is a linear transformation P from a vector space to itself such that

$$P^2 = P$$

That is, whenever P is applied twice to any value, it gives the same result as if it were applied once (**idempotent**). It leaves its image unchanged. Though abstract, this definition of "projection" formalizes and generalizes the idea of graphical projection.

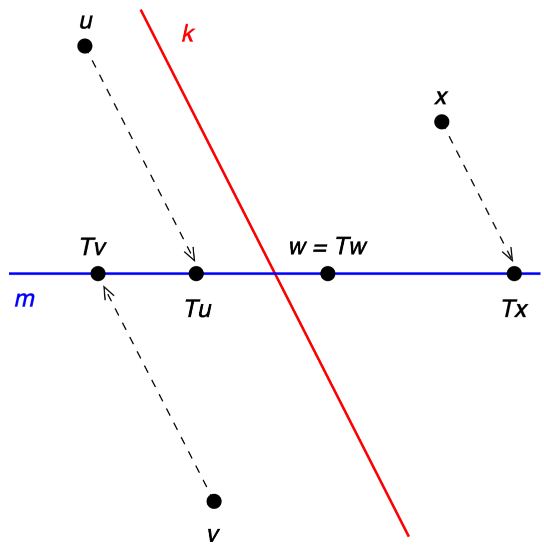

Figure 3: The transformation T is the projection along k onto m. The **image** of T is m and the **null space** is k.

4.1 Properties and classification

Let W be a finite dimensional vector space and P be a projection on W

$$P : W \to U$$

Suppose the subspaces U and V are the **image** and **kernel** of P respectively. Then P has the following properties:

1. By definition, P is idempotent (i.e. $P^2 = P$).

2. P is the identity operator I on U

$$Px = x, \qquad \forall x \in U.$$

3. We have a direct sum $W = U \oplus V$. i.e. Every vector $x \in W$ may be decomposed uniquely as
$$x = u + v$$
with
$$u = Px$$
and
$$v = x - Px = (I - P)x$$
and where $u \in U, v \in V$.

The image and kernel of a projection are complementary, as are P and $Q = I - P$.

The operator Q is also a projection as the image and kernel of P become the kernel and image of Q and vice versa. We say P is a projection along V onto U (kernel/range) and Q is a projection along U onto V.

4.2 Orthogonal projections

An **orthogonal projection** is a projection for which the image U and the null space V are orthogonal subspaces. Thus, for every x and y in W,

$$\langle Px, (y - Py)\rangle = \langle (x - Px), Py\rangle = 0.$$

Equivalently,

$$\langle x, Py\rangle = \langle Px, Py\rangle = \langle Px, y\rangle$$

A projection is orthogonal if and only if it is **self-adjoint**, i.e

$$\langle x, Py \rangle = \langle Px, y \rangle$$

This is because Using the self-adjoint and idempotent properties of P, for any x and y in W, we have $Px \in U$, $y - Py \in V$, and

$$\langle Px, y - Py \rangle = \langle P^2 x, y - Py \rangle = \langle Px, P(I - P)y \rangle = \langle Px, (P - P^2)y \rangle = 0$$

Properties and special cases An orthogonal projection is a bounded operator. This is because for every v in the vector space we have, by Cauchy–Schwarz inequality:

$$\|Pv\|^2 = \langle Pv, Pv \rangle = \langle Pv, v \rangle \leqslant \|Pv\| \cdot \|v\|$$

Thus $\|Pv\| \leqslant \|v\|$.

4.2.1 Formulas

A **simple case** occurs when the orthogonal projection is **onto a line**. If u is a unit vector on the line, i.e. $\|u\| = 1$, then the projection is given by the outer product

$$P_u = uu^{\mathrm{T}}.$$

This operator leaves u invariant, and it annihilates all vectors orthogonal to u, proving that it is indeed the orthogonal projection onto the line containing u. A simple way to see this is to consider an arbitrary vector x as the sum of a component on the line u (i.e. the projected vector we seek) and another perpendicular to u, $x = x_{\|} + x_{\perp}$. Applying projection, we get

$$P_u x = uu^{\mathrm{T}} x_{\|} + uu^{\mathrm{T}} x_{\perp} = u\left(\operatorname{sign}(u^{\mathrm{T}} x_{\|})\|x_{\|}\|\right) + u \cdot 0 = x_{\|},$$

by the properties of the dot product of parallel and perpendicular vectors.

This formula can be generalized to **orthogonal projections on a subspace** of arbitrary dimension. Let $u_1, \ldots, u_k$ be an orthonormal basis of the subspace U, and let A denote the $n \times k$ matrix whose columns are $u_1, \ldots, u_k$. Then the projection is given by:

$$P_A = AA^{\mathrm{T}}$$

which can be rewritten as

$$P_A = \sum_i \langle u_i, \cdot \rangle u_i.$$

The **orthonormality condition can also be dropped**. If $u_1, \ldots, u_k$ is a (not necessarily orthonormal) basis, and A is the matrix with these vectors as columns, then the projection is:

$$P_A = A(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}.$$