GI07/COMPM012: Mathematical Programming and Research Methods (Part 2)

1. Linear Algebra Review

Massimiliano Pontil

Prerequisites & assessment

- Calculus (real-valued functions, limits, derivatives, etc.)

- Fundamentals of linear algebra (vectors, angles, matrices, eigenvectors/eigenvalues,...)

- 1 long homework assignment near the end of the course (35%) – deliver it on-time, penalty otherwise

Material

- Lecture notes
  - http://www.cs.ucl.ac.uk/staff/M.Pontil/courses/index-GI07.htm
- Reference book
  - This and next lecture: Trefethen and Bau. Numerical linear algebra. SIAM.
- Additional material (see web-page for more info)

Course outline

- (Weeks 1,2) Elements of linear algebra and singular value decomposition (SVD)

- (Week 3) Applications of SVD in ML and data analysis

- (Week 4) Elements of graph theory. Applications in ML and data analysis

- (Week 5) Kernel methods

Today's plan

- Linear algebra review
  - vector and matrix operations
  - orthogonality
  - norms
- singular value decomposition

Vectors

- denoted by lower case letters, x, y, b etc.

- they form a *linear space*: 1) $x+y$ is still a vector; 2) If $\lambda \in \mathbb{R}$, λx is still a vector; 3) there is a zero vector, called 0, such that $x+0=x$, etc.

- a vector can be represented by its coefficients relative to a fixed set (basis) of linearly independent vectors $e_1, \ldots, e_n$. The number n is *uniquely* defined as the dimension of the space, which we call $\mathbb{R}^n$

- The coordinate vectors $e_1 = (1, 0, ..., 0)$, $e_2 = (0, 1, 0, ..., 0)$, $\ldots$ $e_n = (0, 0, ..., 0, 1)$ form a basis of $\mathbb{R}^n$ called the standard basis

- x is identified by $(x_1, x_2, ..., x_n)$ since: $(x_1, x_2, ..., x_n) = x_1 e_1 + x_2 e_2 + \ldots x_n e_n$

Matrices

- denoted by upper case letters (A, B etc.). An $m \times n$ matrix is denoted as $A = (A_{ij} : 1 \leq i \leq m, 1 \leq j \leq n)$

- think of a matrix as a *linear transformation* from $\mathbb{R}^n$ to $\mathbb{R}^m$

- they form a linear space (can be viewed as mn-dim vectors)

- denote by a_i the columns of A. Also use the notation $A = [a_1, \ldots, a_n]$

- $Ax = \sum_{i=1}^n x_i a_i$ (linear combination of column vectors)

Matrices (cont.)

- transpose: given $A \in \mathbb{R}^{m\times n}$ its transpose $A^\top \in \mathbb{R}^{n\times m}$ is defined as $A^\top_{ji} = A_{ij}$

- an $n \times n$ matrix is said: *symmetric* if $A_{ij} = A_{ji}$

- *skew symmetric* (or antisymmetric) if $A_{ij} = -A_{ji}$

- positive semi-definite (psd) if $x^\top A x = \sum_{i,j=1}^n x_i A_{ij} x_j \geq 0$ for every $x \in \mathbb{R}^n$ (example: the empirical covariance is symmetric and psd)

Range and null space

- the range space of A is the set of vectors that can be expressed as Ax for some x:

$$\text{range}(A) = \{b : b = Ax, \text{ for some } x \in \mathbb{R}^n\}$$

namely, the set of vectors spanned by the columns of A (so the range of A is also called the column space of A)

- the null space of A is the set of vectors x which satisfy $Ax = 0$:

$$\text{null}(A) = \{x : Ax = 0\}$$

Rank

The column rank of A is the dimension of its columns space

The row rank of A is the dimension of its row space

Theorem: the column rank equals the row rank (we thus refer to this number simply as the rank)

An $m \times n$ matrix A, is said to have *full rank* if rank(A) $=$ min(m, n)

A full rank matrix defines a one-to-one map:

Theorem: An $m \times n$ matrix A, with $m \geq n$ has full rank iff it maps no two distinct vectors to the same vector

Rank one matrices

If A has rank one then $\text{range}(A) = \text{span}\{b\}$, that is

$$Ax = \lambda(x)b$$

by linearity $\lambda(x) = c^\top x$. We arrive to the expression

$$A = bc^\top$$

Two particular cases are

- If $c = e_j$ then all columns of A are zero except the jth column which is c
- If $b = e_i$ then all rows of A are zero except the ith row which equal $c^\top$

Inverse

A square and full rank matrix A is called nonsingular or *invertible*

Since the columns are a basis of $\mathbb{R}^m$, we can write any vector as a *unique* linear combination of them. In particular

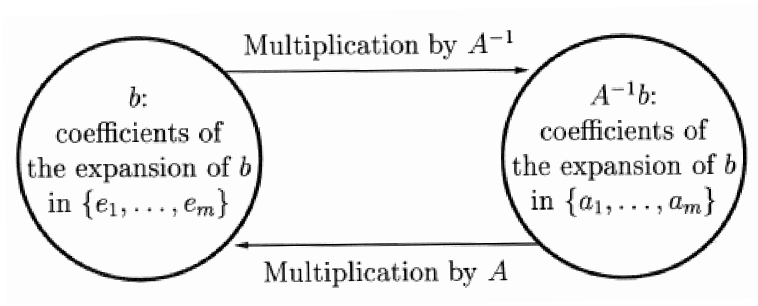
$$e_j = \sum_{i=1}^{m} z_{ij} a_i \quad \text{or } I = AZ$$

Matrix Z is uniquely defined by the above equation. It is called the inverse of A and is denoted as A^{-1}.

Product of invertible matrices: $(AB)^{-1} = B^{-1}A^{-1}$ (analogous to $(AB)^\top = B^\top A^\top$)

Inverse (cont.)

Since $AA^{-1} = A^{-1}A = I$, the equation $Ax = b$ has always a unique solution, given by $A^{-1}b$. Interpretation: think of $A^{-1}b$ as the vector of coefficients of the expansion of b in the basis of columns of A



$$Ax = b \iff Ax = AA^{-1}b \iff x = A^{-1}AA^{-1}b = A^{-1}b$$

Orthogonal vectors

Recall the notion of **inner product:** $x^\top y = \sum_{i=1}^n x_i y_i$

and **Euclidean norm:** $\|x\| = \sqrt{x^\top x}$

A pair of vectors x and y are called orthogonal if $x^\top y = 0$

The set $S = \{u_1, \ldots, u_k\}$ is called orthogonal if its elements are pairwise orthogonal; if, in addition, $\|u_i\| = 1$ for $i = 1, \ldots, k$ then S is said orthonormal

Theorem: the vectors in an orthogonal set $\{u_1, \ldots, u_k\}$ are linearly independent

Proof (hint) assume by contradiction that u_1 is a linear combination of $u_2, \ldots, u_m$ and conclude that $u_1 = 0$

Orthogonal vectors (cont.)

If $S = \{u_1, \ldots, u_k\}$ is an orthonormal (o.n.) set and x an arbitrary vector in $\mathbb{R}^m$, the vector

$$r = x - \sum_{i=1}^{k} (u_i^\top x) u_i$$

is orthogonal to S.

In particular, if $k = m$, then S is a basis and r must be zero

The linear space $\{y : u_i^\top y = 0, i = 1, \ldots, k\}$ is called the orthogonal complement to S

Orthogonal matrices

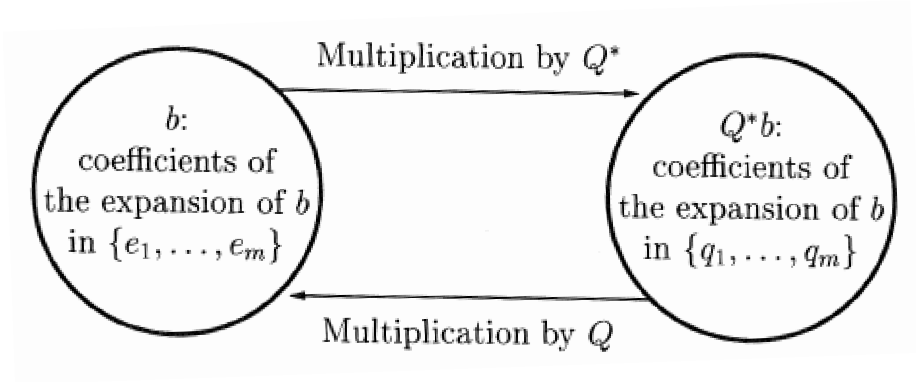
If $\{u_1, \ldots, u_k\}$ is an o.n. set then the $m \times k$ matrix $U = [u_1, \ldots, u_k]$ has the property that $U^\top U = I_{k\times k}$

When $k = m$ the matrix U is said orthogonal. In this case we have that $U^{-1} = U^\top$, that is

$$U^\top U = I_{m\times m} \qquad \text{(or equivalently } UU^\top = I_{m\times m}\text{)}$$

Orthogonal matrices (cont.)

Interpretation:



Note that the transformation U preserves the inner product (so the angles and lengths of vectors are preserved)

$$(Ux)^\top (Uy) = x^\top y$$

If $\det(U) = 1$ then U is a rotation; if $\det(U) = -1$ then U is a reflection

Norms

A norm is a function $\|\cdot\| : \mathbb{R}^m \to [0,\infty)$ which measures the length of a vector. It satisfies the conditions

- $\|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$
- $\|x+y\| \leq \|x\| + \|y\|$ (triangle inequality)
- $\|\alpha x\| = |\alpha|\|x\|$

for all $x, y \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$

Norms (cont.)

Norms are convex: for all $\lambda \in [0, 1]$, $x, y \in \mathbb{R}^m$ we have

$$\|\lambda x + (1 - \lambda)y\| \leq \lambda\|x\| + (1 - \lambda)\|y\|$$

An important class of norms are the p-norms:

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p\right)^{1/p}, \quad \text{for } p \geq 1$$

and

$$\|x\|_\infty = \max_{i=1}^{m} |x_i|$$

Induced matrix norms

The space of $m \times n$ matrices is an mn-dimensional space. Any norm on this space can be used to define the size of such matrices

An induced matrix norm is a special type of norm associated with matrices, which is induced by the norms in the domain and codomain of A:

$$\|A\|_{(m,n)} = \sup_{x \in \mathbb{R}^n} \frac{\|Ax\|_{(m)}}{\|x\|_{(n)}}$$

(can you argue this is a norm?)

Induced matrix norms (cont.)

For example if $\|x\|_{(n)}$ and $\|Ax\|_{(m)}$ are the standard Euclidean norms

$$\|A\| = \sup_{x \in \mathbb{R}^n} \sqrt{\frac{x^\top A^\top A x}{x^\top x}} = \sqrt{\lambda_{\mathsf{max}}(A^\top A)}$$

An important property of induced matrix norms is:

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

This follows by $\|Ax\|_{(m)} \leq \|A\|\|x\|_{(n)}$

Frobenius norm

An important example of matrix norms which is not induced by vector norms is the Frobenius norm

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

This is the standard Euclidean norm when matrix A is viewed as an mn-dimensional vector. It may also be written as

$$\|A\|_F = \left(\sum_{j=1}^{n} \|a_j\|_2^2 \right)^{1/2}$$

or as

$$\|A\|_F = \sqrt{\text{trace}(A^\top A)} = \sqrt{\text{trace}(AA^\top)}$$

(the trace of a matrix is the sum of the diagonal elements)

Singular value decomposition (SVD)

- SVD is a matrix factorization whose computation is key in many algorithms

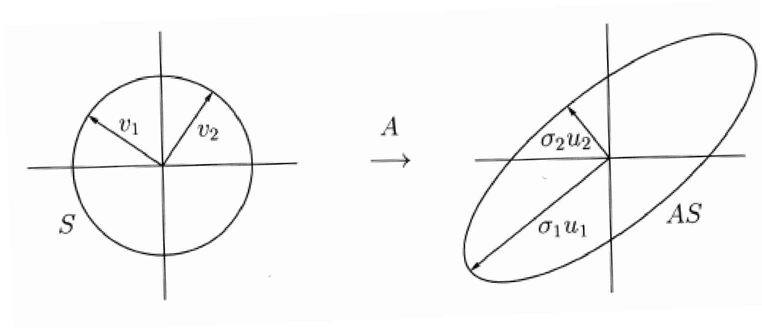
- many ML and statistical methods are based on SVD:

  - least squares, regularization
  - principal component analysis
  - spectral clustering
  - matrix factorization, etc.

- being familiar with SVD is essential in order to understand and implement ML/statistical methods

What is it?

Observation: the image of the unit hypersphere under any $m \times n$ matrix A is an hyperellipse



Hyperellipse: surface in $\mathbb{R}^m$ obtained by stretching the unit sphere in $\mathbb{R}^m$ by some nonnegative factors $\sigma_1, \ldots, \sigma_m$ in some orthogonal directions (unit vectors) $u_1, \ldots, u_m$

The vectors $\{\sigma_i u_i\}$ are the principal axes of the hyperellipse, with lengths $\sigma_1, \ldots, \sigma_m$ (use the convention that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$)

What is it? (cont.)

- we call the singular values of A the lengths of the n principal axis of AS,
- the left singular vectors of A, $u_1, \ldots, u_n$, are the principal semiaxes of AS
- the right singular vectors of A, $v_1, \ldots, v_n$, are the preimages of the principal semiaxis of AS
- if $m \geq n$ at most n of the σ_i are nonzero
- if A has rank r, exactly r of the σ_i are nonzero

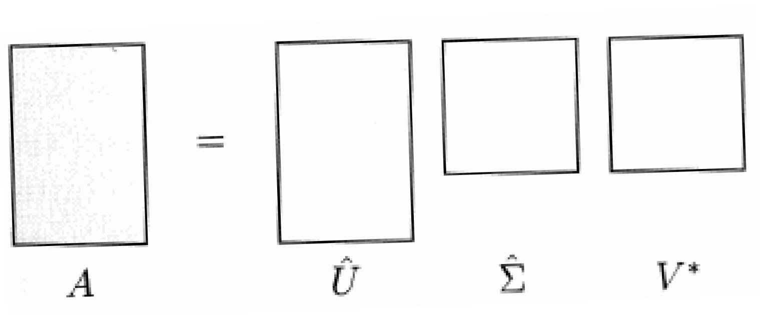
Reduced SVD

Assume for simplicity that rank$(A) = n$. We have seen that

$$Av_j = \sigma_j u_j, \quad j \in \{1, \ldots, n\}$$

or, $AV = \hat{U}\hat{\Sigma}$, with $\hat{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, $\hat{U} = [u_1, \ldots, u_n]$ and V is an $n \times n$ orthogonal matrix. We may then write

$$A = \hat{U}\hat{\Sigma}V^\top$$

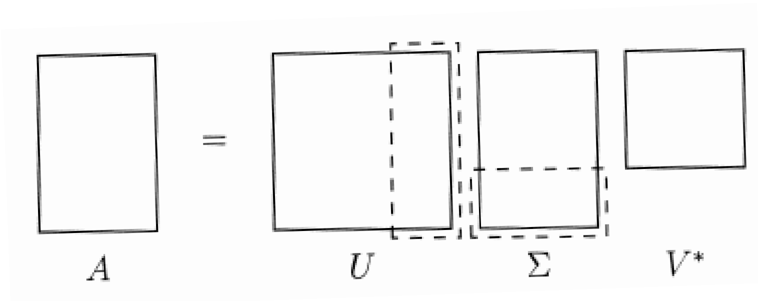


Full SVD

Recall that we assumed $m \geq n$. If $m > n$, we can complete the set $\{u_1, \ldots, u_n\}$ to a basis of $\mathbb{R}^m$ by adding to it $m-n$ additional onthonormal vectors $u_{n+1}, \ldots, u_m$.

We replace $\hat{U}$ by the orthogonal matrix $U = [u_1, \ldots, u_m]$ and $\hat{\Sigma}$ by the $m \times n$ matrix Σ having $\hat{\Sigma}$ in the upper $n \times n$ block and $m-n$ zero rows below it. This gives us a new factorization of A

$$A = U\Sigma V^\top$$



Formal definition

Given an $m \times n$ real matrix A, a singular value decomposition (SVD) of A is a factorization

$$A = U \Sigma V^\top$$

where: U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix and Σ is diagonal

Also use the convention that the diagonal entries of Σ are nonnegative and nonincreasing:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0, \quad p = \min(n, m)$$

Existence and uniqueness

Theorem: every $m \times n$ matrix A has an SVD, whose singular values σ_j are uniquely determined. Moreover, if $m = n$ and the singular values are distinct, the left and right singular vectors are uniquely determined up to a sign change

Proof idea is to isolate the direction of the largest action of A and then proceed by induction

Change of basis

Another interpretation of SVD: every matrix is diagonal if one uses the proper bases for the domain and range spaces

$$b = Ax \iff U^\top b = U^\top Ax = U^\top U\Sigma V^\top x \iff b' = \Sigma x'$$

where $b' = U^\top b$ and $x' = V^\top x$

- range space is expressed in the basis of columns of U

- domain space is expressed in the basis of columns of V

Properties of SVD

- if A is a rank one matrix, $A = bc^\top$, we have $\sigma_1 = \|b\|\|c\|$ and $u_1 = \frac{b}{\|b\|}$, $v_1 = \frac{c}{\|c\|}$ (up to a sign change)

- the rank r of a matrix A equals the number of nonzero singular values

  Proof: $A = U\Sigma V^\top$. Now the rank of Σ is r. Since U and V are full rank, it follows that rank(A) = rank(Σ)

- range(A) = span$\{u_1, \ldots, u_r\}$; null(A) = span$\{v_{r+1}, \ldots, v_n\}$

Properties of SVD (cont.)

- $\sigma_1 = \|A\|_{(2,2)}$

- The nonzero singular values of A are the square root of the nonzero eigenvalues of $A^\top A$ or $AA^\top$

- If A is a square symmetric matrix, then the nonzero singular values of A are the absolute value of the eigenvalues of A

Low rank approximation

Another way to explain the SVD is to see A as a sum of rank one matrices

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^\top \qquad (*)$$

There are many ways to express A as sum or rank matrices (can you think of any?). Formula $(*)$ has however a special property (which, as we will see later is important e.g. in PCA).

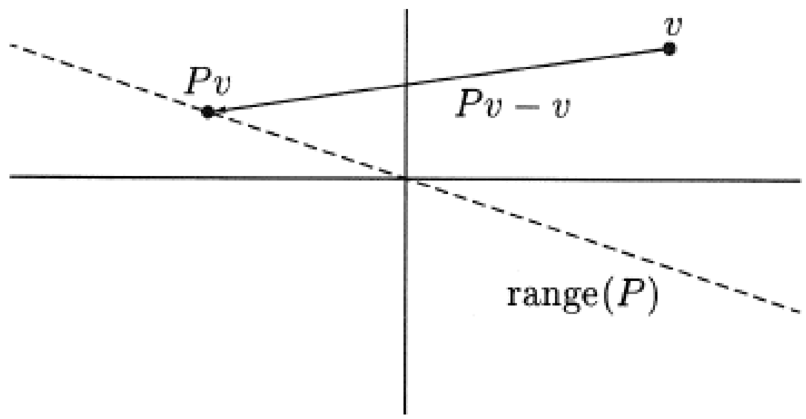
Let $k \le r$. We will see that the k-th partial sum, $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^\top$, captures much of the "energy" of A as possible:

$$\|A - A_k\|_{2,2} = \min\{\|A - B\|_{2,2} : B \in \mathbb{R}^{m \times n}, \text{rank}(B) \le k\}$$

Projection

A projection is a square matrix P such that $P^2 = P$

For every v we have that $Pv - v$ is in the null space of P because $P(Pv - v) = (P^2 - P)v = 0$



Complementary projection

If P is a projection, $I - P$ is also a projection:

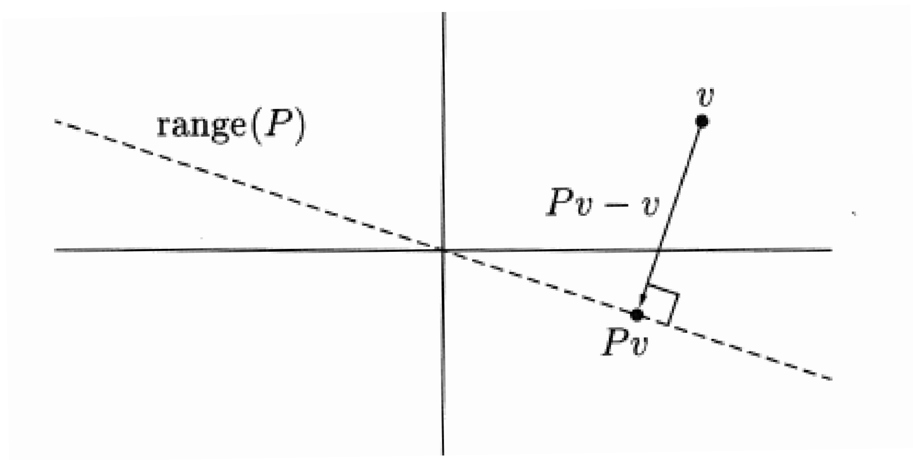
$$(I-P)^2 = I^2 + P^2 - 2IP = I + P - 2P = I - P$$

Moreover, range$(I - P)$ = null(P) because $P((I - P)v) = 0$. Likewise, range(P) = null$(I - P)$

Since range$(P) \cap$ null$(P) = \{0\}$ we see that a projection separates $\mathbb{R}^n$ into two spaces

Orthogonal projections

An orthogonal projection is one such that range(P) is orthogonal to null(P).



Theorem: A projection P is orthogonal iff P is symmetric

Orthogonal projections (cont.)

An orthogonal projection is expressed as

$$P = \widehat{U}\widehat{U}^\top = \sum_{i=1}^{k} u_i u_i^\top$$

where $\widehat{U} = [u_1, \ldots, u_k]$ and the u_i are o.n. vectors

If $u_{k+1}, \ldots, u_n$ complete the set $\{u_1, \ldots, u_k\}$ to an o.n. basis, the orthogonal projection $I - P$ can be written as

$$\sum_{i=k+1}^{n} u_i u_i^\top$$