

Lecture Note: Matrix Operations and Matrix Decompositions

Focus: algebraic structure, computational procedures, theoretical properties, and statistical applications

Why matrix operations matter

- ▶ Linear models reduce to matrix equations.
- ▶ Covariance matrices encode dependence and uncertainty.
- ▶ Optimization algorithms repeatedly solve linear systems.
- ▶ PCA, latent-factor models, and low-rank learning are spectral problems.
- ▶ Numerical stability is often more important than a closed-form inverse.

Recurring computational tasks

Given a matrix A , we often want to:

- ▶ solve $Ax = b$,
- ▶ compute quadratic forms,
- ▶ identify rank and null spaces,
- ▶ invert or partially invert blocks,
- ▶ expose dominant directions.

Notation and dimensions

Let

$$A \in \mathbb{R}^{m \times n}, \quad x \in \mathbb{R}^n, \quad b \in \mathbb{R}^m.$$

- ▶ A_{ij} or a_{ij} : entry in row i , column j .
- ▶ $A_{i:}$: row i ; $A_{:j}$: column j .
- ▶ A^T : transpose; A^{-1} : inverse when it exists.
- ▶ I_n : $n \times n$ identity matrix.
- ▶ For symmetric A , $A = A^T$.

Dimension check

If $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, then $AB \in \mathbb{R}^{m \times p}$. Dimension compatibility is the first debugging tool in matrix algebra.

Matrix-vector and matrix-matrix multiplication

For $y = Ax$,

$$y_i = \sum_{j=1}^n a_{ij}x_j.$$

Equivalently,

$$Ax = \sum_{j=1}^n x_j A_{:j},$$

so multiplication forms a linear combination of the columns of A .

For $C = AB$,

$$c_{ij} = A_{i:} B_{:j} = \sum_{k=1}^n a_{ik} b_{kj}.$$

Key point

Matrix multiplication is associative, but generally not commutative:

$$A(BC) = (AB)C, \quad AB \neq BA \text{ in general.}$$

Transpose, inverse, and solving systems

Useful identities:

$$(AB)^T = B^T A^T, \quad (A^{-1})^T = (A^T)^{-1}, \quad (AB)^{-1} = B^{-1} A^{-1}.$$

If A is square and nonsingular, the formal solution of $Ax = b$ is

$$x = A^{-1}b.$$

Special matrix classes

- ▶ **Diagonal:** $A = \text{diag}(d_1, \dots, d_n)$.
- ▶ **Triangular:** entries vanish above or below the diagonal.
- ▶ **Symmetric:** $A = A^T$.
- ▶ **Orthogonal:** $Q^T Q = Q Q^T = I$.
- ▶ **Projection:** $P^2 = P$; orthogonal if also $P = P^T$.
- ▶ **Positive definite:** $x^T A x > 0$ for every $x \neq 0$.
- ▶ **Positive semidefinite:** $x^T A x \geq 0$.

These structures determine which decomposition is appropriate and which algorithms are stable.

Rank, column space, and null space

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

$$\text{col}(A) = \{Ax : x \in \mathbb{R}^n\}, \quad \text{null}(A) = \{x : Ax = 0\}.$$

The rank is

$$\text{rank}(A) = \dim\{\text{col}(A)\}.$$

The rank-nullity theorem gives

$$\text{rank}(A) + \dim\{\text{null}(A)\} = n.$$

Statistical interpretation

For a design matrix X , full column rank means the regression coefficients are identifiable from $X^T X$. Rank deficiency signals exact linear dependence among predictors.

Trace, determinant, and norms

- ▶ Trace: $\text{tr}(A) = \sum_i a_{ii}$ and $\text{tr}(AB) = \text{tr}(BA)$ when products conform.
- ▶ Determinant: $\det(AB) = \det(A) \det(B)$.
- ▶ Frobenius norm:

$$\|A\|_F = \left(\sum_{i,j} a_{ij}^2 \right)^{1/2} = \sqrt{\text{tr}(A^T A)}.$$

- ▶ Spectral norm:

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

Later we will see

$$\|A\|_2 = \sigma_{\max}(A), \quad \|A\|_F^2 = \sum_i \sigma_i^2.$$

Positive definiteness

For symmetric $A \in \mathbb{R}^{n \times n}$, the following are equivalent:

- ① $x^T A x > 0$ for all $x \neq 0$;
- ② all eigenvalues of A are positive;
- ③ all leading principal minors are positive;
- ④ A has a Cholesky factorization $A = LL^T$ with positive diagonal entries.

Examples

Covariance matrices are positive semidefinite. If no nonzero linear combination of variables has zero variance, the covariance matrix is positive definite.

Block matrices and the Schur complement

Partition a symmetric matrix as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad A_{11} \text{ nonsingular.}$$

The Schur complement of A_{11} is

$$S = A_{22} - A_{21}A_{11}^{-1}A_{12}.$$

- ▶ Appears in block Gaussian elimination.
- ▶ Appears after sweeping a block of variables.
- ▶ Determines conditional covariance in Gaussian models.
- ▶ Helps characterize positive definiteness.

If $A_{11} \succ 0$, then $A \succ 0$ iff $S \succ 0$.

Numerical themes: cost, stability, conditioning

Three questions should accompany every matrix formula:

- ① **Cost:** How many floating-point operations are needed?
- ② **Stability:** Does rounding error get amplified by the algorithm?
- ③ **Conditioning:** Is the mathematical problem itself sensitive to perturbations?

For nonsingular square A ,

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}.$$

A large condition number means small changes in data can cause large changes in the solution.

The SWEEP operator: motivation

The SWEEP operator is a compact elimination procedure for symmetric matrices. It is especially useful when the matrix contains cross-products such as

$$\begin{pmatrix} X^T X & X^T y \\ y^T X & y^T y \end{pmatrix}.$$

Sweeping predictor indices can reveal, in one transformed matrix:

- ▶ regression coefficients,
- ▶ residual sums of squares,
- ▶ blocks related to $(X^T X)^{-1}$,
- ▶ partial and sequential regression information.

Definition: sweep on one pivot

Let $A = (a_{ij})$ be symmetric and suppose $a_{kk} \neq 0$. Define $B = \text{SWP}_k(A)$ by

$$\begin{aligned}b_{kk} &= -\frac{1}{a_{kk}}, \\b_{ik} &= \frac{a_{ik}}{a_{kk}}, \quad i \neq k, \\b_{kj} &= \frac{a_{kj}}{a_{kk}}, \quad j \neq k, \\b_{ij} &= a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}}, \quad i, j \neq k.\end{aligned}$$

The last line is a rank-one Schur-complement update.

Convention

Different texts use slightly different sign conventions. This lecture uses the symmetric convention above.

Algebraic interpretation of one sweep

Write the pivoted matrix, after reordering if necessary, as

$$A = \begin{pmatrix} a & r^T \\ r & C \end{pmatrix}.$$

Sweeping the first index gives

$$\text{SWP}_1(A) = \begin{pmatrix} -a^{-1} & a^{-1}r^T \\ a^{-1}r & C - ra^{-1}r^T \end{pmatrix}.$$

Interpretation

The unswept block becomes the Schur complement $C - ra^{-1}r^T$. Thus SWEEP combines elimination, partial inversion, and Schur complementation.

Connection to Gaussian elimination

Suppose

$$\begin{pmatrix} a & r^T \\ r & C \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Eliminating x_1 produces

$$(C - ra^{-1}r^T)x_2 = b_2 - ra^{-1}b_1.$$

The same Schur-complement update appears in a sweep.

- ▶ Gaussian elimination targets a linear system.
- ▶ SWEEP stores elimination information inside a symmetric matrix.
- ▶ Repeated sweeps are therefore closely related to symmetric elimination.

Worked example: one sweep

Consider

$$A = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 6 \end{pmatrix}.$$

Sweep index 1 using pivot $a_{11} = 4$:

$$\text{SWP}_1(A) = \begin{pmatrix} -1/4 & 1/2 & 1/4 \\ 1/2 & 4 & 5/2 \\ 1/4 & 5/2 & 23/4 \end{pmatrix}.$$

For example,

$$b_{23} = 3 - \frac{2 \cdot 1}{4} = \frac{5}{2}.$$

The lower-right 2×2 block is the Schur complement of the pivot.

Complete sweep and matrix inversion

Let $K = \{k_1, \dots, k_q\}$ be a set of indices. Repeated sweeping gives

$$\text{SWP}_K(A) = \text{SWP}_{k_q} \cdots \text{SWP}_{k_1}(A).$$

If A is nonsingular and all indices can be swept, then under our convention

$$\text{SWP}_{\{1, \dots, n\}}(A) = -A^{-1}.$$

For a 2×2 matrix

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

complete sweeping yields

$$-\frac{1}{ac - b^2} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix} = -A^{-1}.$$

This is one reason SWEEP is useful for updating inverse-like information without explicitly recomputing an inverse from scratch.

Regression cross-product matrix

For the linear model

$$y = X\beta + \varepsilon,$$

form the augmented cross-product matrix

$$A = \begin{pmatrix} X^T X & X^T y \\ y^T X & y^T y \end{pmatrix}.$$

If X has p columns and we sweep the first p indices, then

$$\text{SWP}_{1:p}(A) = \begin{pmatrix} -(X^T X)^{-1} & \hat{\beta} \\ \hat{\beta}^T & \text{SSE} \end{pmatrix},$$

where

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Reading regression quantities from the swept matrix

After sweeping all predictor indices:

- ▶ last column, predictor rows: $\hat{\beta}$;
- ▶ last diagonal entry: residual sum of squares;
- ▶ swept predictor block: $-(X^T X)^{-1}$.

With $\hat{\sigma}^2 = \text{SSE} / (n - p)$,

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}.$$

Therefore the diagonal of the negative swept predictor block gives the variance multipliers for coefficient standard errors.

Worked regression example with SWEEP

Take

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}.$$

Then

$$A = \begin{pmatrix} 3 & 3 & 5 \\ 3 & 5 & 6 \\ 5 & 6 & 9 \end{pmatrix}.$$

Sweeping the two predictor indices gives approximately

$$\begin{pmatrix} -0.8333 & 0.5000 & 1.1667 \\ 0.5000 & -0.5000 & 0.5000 \\ 1.1667 & 0.5000 & 0.1667 \end{pmatrix}.$$

Hence $\hat{\beta} = (1.1667, 0.5000)^T$ and $SSE = 0.1667$.

Cholesky decomposition

For symmetric positive-definite A ,

$$A = LL^T,$$

where L is lower triangular with positive diagonal entries. Equivalently,

$$A = R^T R,$$

with $R = L^T$ upper triangular.

Why it matters

Cholesky is the preferred dense factorization for symmetric positive definite (SPD) systems: it exploits symmetry, needs about half the storage of a general factorization, and is faster than generic LU.

Deriving the Cholesky recursion

From $A = LL^T$,

$$a_{ij} = \sum_{k=1}^{\min(i,j)} \ell_{ik} \ell_{jk}.$$

For diagonal elements,

$$\ell_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} \ell_{jk}^2}.$$

For $i > j$,

$$\ell_{ij} = \frac{1}{\ell_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} \ell_{jk} \right).$$

Each column is computed from previously completed columns.

Solving $Ax = b$ using Cholesky

Given $A = LL^T$, solve

$$LL^T x = b$$

in two triangular stages:

- ① Forward substitution: solve $Lz = b$.
- ② Back substitution: solve $L^T x = z$.

Each triangular solve costs $O(n^2)$ after the $O(n^3)$ factorization.

Multiple right-hand sides

For $AX = B$, factor A once and solve for every column of B using the same L .

Determinants and log-determinants

If $A = LL^T$, then

$$\det(A) = \det(L)^2 = \left(\prod_{i=1}^n \ell_{ii} \right)^2.$$

Therefore

$$\log \det(A) = 2 \sum_{i=1}^n \log \ell_{ii}.$$

Why this is important

Gaussian likelihoods repeatedly require

$$-\frac{1}{2} \log \det(\Sigma) - \frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu).$$

Cholesky evaluates both terms stably without forming Σ^{-1} explicitly.

Quadratic forms and Gaussian simulation

To compute

$$q = b^T A^{-1} b,$$

solve $Lz = b$. Since $A^{-1} = L^{-T} L^{-1}$,

$$q = z^T z = \|z\|_2^2.$$

For simulation, if $z \sim N(0, I)$ and $A = LL^T$, then

$$x = \mu + Lz \sim N(\mu, A).$$

Thus the same factorization supports solving, likelihood evaluation, quadratic forms, and Gaussian sampling.

Cholesky and least squares

The normal equations are

$$X^T X \hat{\beta} = X^T y.$$

If X has full column rank, then $X^T X \succ 0$ and can be factored:

$$X^T X = LL^T.$$

Then solve

$$Lz = X^T y, \quad L^T \hat{\beta} = z.$$

Caution

Forming $X^T X$ squares the 2-norm condition number:

$$\kappa_2(X^T X) = \kappa_2(X)^2.$$

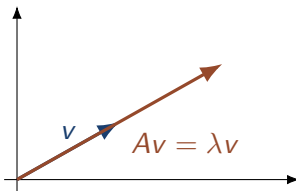
For ill-conditioned problems, QR or SVD is preferable.

Eigenvalues and eigenvectors

For a square matrix A , a nonzero vector v is an eigenvector with eigenvalue λ if

$$Av = \lambda v.$$

Geometrically, A maps the direction v to itself, changing only its scale and possibly its sign.



Characteristic polynomial and spectrum

An eigenvalue satisfies

$$(A - \lambda I)v = 0$$

for some $v \neq 0$, hence

$$\det(A - \lambda I) = 0.$$

The roots of the characteristic polynomial are the eigenvalues.

- ▶ The set of eigenvalues is the **spectrum** of A .
- ▶ $\text{tr}(A) = \sum_i \lambda_i$ counting algebraic multiplicity.
- ▶ $\det(A) = \prod_i \lambda_i$.

Computation

Numerical eigensolvers do *not* usually form and solve the characteristic polynomial; that route is unstable and inefficient.

Spectral theorem for symmetric matrices

Theorem

If $A = A^T \in \mathbb{R}^{n \times n}$, then all eigenvalues are real and there exists an orthogonal matrix Q such that

$$A = Q\Lambda Q^T,$$

where

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n).$$

Consequences:

- ▶ eigenvectors associated with distinct eigenvalues are orthogonal;
- ▶ $Q^{-1} = Q^T$;
- ▶ symmetric matrices have a complete orthonormal eigenbasis.

Worked eigendecomposition

Consider

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Its eigenpairs are

$$\lambda_1 = 3, \quad q_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_2 = 1, \quad q_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Thus

$$A = Q \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} Q^T, \quad Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The transformation stretches most strongly along the $(1, 1)^T$ direction.

Quadratic forms and definiteness via eigenvalues

If $A = Q\Lambda Q^T$ and $z = Q^T x$, then

$$x^T A x = z^T \Lambda z = \sum_{i=1}^n \lambda_i z_i^2.$$

Therefore:

- ▶ $A \succ 0$ iff $\lambda_i > 0$ for every i ;
- ▶ $A \succeq 0$ iff $\lambda_i \geq 0$ for every i ;
- ▶ A is indefinite if it has both positive and negative eigenvalues.

This diagonal-coordinate view is central to optimization and covariance analysis.

Matrix powers and functions

For a diagonalizable symmetric matrix $A = Q\Lambda Q^T$,

$$A^k = Q\Lambda^k Q^T.$$

More generally, define

$$f(A) = Qf(\Lambda)Q^T,$$

where

$$f(\Lambda) = \text{diag}(f(\lambda_1), \dots, f(\lambda_n)).$$

Examples:

$$A^{1/2} = Q\Lambda^{1/2}Q^T, \quad \exp(A) = Q \exp(\Lambda)Q^T.$$

For $A \succ 0$, this also defines $A^{-1/2}$, useful for whitening and standardization.

Rayleigh quotient and extremal eigenvalues

For symmetric A , define the Rayleigh quotient

$$R_A(x) = \frac{x^T A x}{x^T x}, \quad x \neq 0.$$

The variational characterization is

$$\lambda_{\max}(A) = \max_{\|x\|=1} x^T A x, \quad \lambda_{\min}(A) = \min_{\|x\|=1} x^T A x.$$

Interpretation

Eigenvectors are stationary directions of the quadratic form; extreme eigenvectors identify the directions of largest and smallest curvature.

PCA as an eigenvalue problem

For centered data matrix $X \in \mathbb{R}^{n \times p}$, the sample covariance matrix is

$$S = \frac{1}{n-1} X^T X.$$

The first principal loading solves

$$\max_{\|v\|=1} v^T S v.$$

By the Rayleigh principle, the solution is the eigenvector associated with $\lambda_{\max}(S)$.

- ▶ Eigenvalues measure explained variance.
- ▶ Eigenvectors give orthogonal loading directions.
- ▶ Scores are Xv_j .

Singular value decomposition: statement

For any $A \in \mathbb{R}^{m \times n}$, there exist orthogonal matrices U and V such that

$$A = U\Sigma V^T,$$

where Σ is rectangular diagonal with nonnegative entries

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

The σ_i are the singular values and $r = \text{rank}(A)$.

Key advantage

Unlike an eigendecomposition, the SVD exists for every real rectangular matrix.

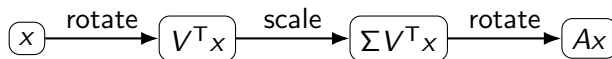
Geometric interpretation of the SVD

The factorization

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

can be read from right to left:

- 1 V^T : rotate or reflect the input coordinates;
- 2 Σ : stretch by factors σ_i ;
- 3 U : rotate or reflect into the output coordinates.



SVD from eigenvalue problems

From

$$A = U\Sigma V^T,$$

we obtain

$$A^T A = V\Sigma^T \Sigma V^T, \quad AA^T = U\Sigma \Sigma^T U^T.$$

Therefore:

- ▶ right singular vectors v_i are eigenvectors of $A^T A$;
- ▶ left singular vectors u_i are eigenvectors of AA^T ;
- ▶ $\sigma_i = \sqrt{\lambda_i(A^T A)}$.

Numerical caution

This identity proves existence, but forming $A^T A$ explicitly can square the condition number. High-quality SVD routines avoid doing so.

Thin/economy SVD and rank structure

If A has rank r , the compact SVD is

$$A = U_r \Sigma_r V_r^T,$$

where

$$U_r \in \mathbb{R}^{m \times r}, \quad \Sigma_r \in \mathbb{R}^{r \times r}, \quad V_r \in \mathbb{R}^{n \times r}.$$

Then

$$\text{col}(A) = \text{span}(u_1, \dots, u_r),$$

$$\text{col}(A^T) = \text{span}(v_1, \dots, v_r),$$

and the remaining right singular vectors span $\text{null}(A)$. The SVD makes the four fundamental subspaces explicit.

Moore-Penrose pseudoinverse

For the compact SVD

$$A = U_r \Sigma_r V_r^T,$$

define

$$A^+ = V_r \Sigma_r^{-1} U_r^T.$$

The pseudoinverse satisfies

$$AA^+A = A, \quad A^+AA^+ = A^+,$$

with AA^+ and A^+A symmetric.

Interpretation

AA^+ is the orthogonal projector onto $\text{col}(A)$, while A^+A projects onto $\text{col}(A^T)$.

Least squares through the SVD

For

$$\min_x \|Ax - b\|_2,$$

the minimum-norm solution is

$$\hat{x} = A^+ b = \sum_{i=1}^r \frac{u_i^T b}{\sigma_i} v_i.$$

This formula exposes instability: components associated with tiny σ_i are amplified by $1/\sigma_i$.

Truncated regularization

Discarding very small singular values gives a truncated-SVD estimator and stabilizes ill-posed problems.

Best low-rank approximation: Eckart-Young-Mirsky

Write

$$A = \sum_{i=1}^r \sigma_i u_i v_i^T.$$

The best rank- k approximation is

$$A_k = \sum_{i=1}^k \sigma_i u_i v_i^T.$$

It solves

$$\min_{\text{rank}(B) \leq k} \|A - B\|_2 = \sigma_{k+1},$$

and

$$\min_{\text{rank}(B) \leq k} \|A - B\|_F = \left(\sum_{i>k} \sigma_i^2 \right)^{1/2}.$$

This theorem is the mathematical foundation of low-rank compression and latent-factor approximation.

PCA computed directly by SVD

For centered $X \in \mathbb{R}^{n \times p}$, let

$$X = U\Sigma V^T.$$

Then

$$S = \frac{1}{n-1} X^T X = V \frac{\Sigma^2}{n-1} V^T.$$

Therefore:

- ▶ PCA loading vectors are columns of V ;
- ▶ explained variances are $\sigma_j^2/(n-1)$;
- ▶ principal component scores are $XV = U\Sigma$.

Computing PCA by SVD of centered data is often preferable to explicitly forming the covariance matrix.

Singular values and conditioning

For full-column-rank A ,

$$\kappa_2(A) = \frac{\sigma_1}{\sigma_n}.$$

A small σ_n means some direction v_n is nearly mapped to zero:

$$Av_n = \sigma_n u_n.$$

Hence:

- ▶ near rank deficiency corresponds to tiny singular values;
- ▶ the SVD is the most transparent tool for diagnosing ill-conditioning;
- ▶ numerical rank can be defined by thresholding singular values relative to σ_1 and machine precision.

Synthesis: which tool should I use?

Tool	Matrix class	Main output	Typical use
SWEEP	symmetric cross-product/block matrix	partial inverse + Schur complements	regression updating, nested models
Cholesky	symmetric positive definite	$A = LL^T$	solves, log-determinants, Gaussian models
Eigen	square; strongest theory for symmetric	$A = Q\Lambda Q^T$	quadratic forms, dynamics, covariance PCA
SVD	any $m \times n$ matrix	$A = U\Sigma V^T$	least squares, rank, PCA, compression, conditioning

Final principle

Exploit matrix structure. Avoid explicit inverses unless the inverse itself is the object of interest. Use orthogonal or triangular factorizations to obtain stable computations.