smfd5.tex
Day 5. 25.10.2013.

5. **Multiplication Theorem for Determinants**.

If A, B are $n \times n$ (so AB, and BA, are defined),

$$|AB| = |A|.|B|.$$

Proof. We can display a matrix A as a row of its columns, $A = [\mathbf{a}_1, \dots, \mathbf{a}_n]$ (or as a column of its rows). The kth column of the matrix product $C = AB$ is then

$$\mathbf{c}_k = b_{1k}\mathbf{a}_1 + \dots + b_{nk}\mathbf{a}_n.$$

For, the ith element of the kth column of C is

$$c_{ik} = \sum_j a_{ij}b_{jk} = \sum_j b_{jk}[\mathbf{a}_j]_i = [\sum_j b_{jk}\mathbf{a}_j]_i.$$

This is the ith element of the above vector equation, on both sides. Then

$$detC = detAB = det[b_{11}\mathbf{a}_1 + \dots + b_{n1}\mathbf{a}_n, \dots, b_{1n}\mathbf{a}_1 + \dots + b_{nn}\mathbf{a}_n].$$

Expand the RHS by the first column. We get a sum of the form

$$\sum_{j_1} b_{j_1,1} det[\dots].$$

Expand each det here by the second column. We get a double sum, of the form

$$\sum_{j_1,j_2} b_{j_1,1}b_{j_2,1} det[\dots],$$

and so on, finally getting

$$\sum_{j_1,\dots,j_n} b_{j_1,1} \dots b_{j_n,1} det[\dots].$$

Each matrix whose det we are taking here is a row of columns of A. Each such det with two columns the *same* vanishes. So we can reduce the 'big' sum (n^n terms) to a smaller sum with all columns *different* ($n!$ terms). Then we have a *permutation* of the columns, σ say, giving

$$detC = \sum_\sigma b_{\sigma(1),1} \dots b_{\sigma(n),n} det[\mathbf{a}_{\sigma(1)}, \dots, \mathbf{a}_{\sigma(n)}].$$

Putting the columns here in their natural order,

$$detC = \sum_{\sigma} b_{\sigma(1),1} \dots b_{\sigma(n),n}.(-1)^{sgn(\sigma)} det[\mathbf{a}_1, \dots, \mathbf{a}_n].$$

The determinant here is $detA$, so we can take it out. This leaves $detB$, so

$$detC = det(AB) = detA.detB. \qquad //$$

6. *Inverses again.*

If A is $n \times n$, the (i,j) *minor* is the determinant of the $(n-1) \times (n-1)$ submatrix obtained by deleting the ith row and jth column. The (i,j) *cofactor*, or *signed minor* A_{ij}, is the (i,j) minor times $(-)^{i+j}$ (the signs follow a chessboard or chequerboard pattern, with + in the top left-hand corner),

The matrix $B = (b_{ij})$, where

$$b_{ij} := A_{ji}/|A|,$$

is the *inverse matrix* A^{-1} of A, defined iff $|A| \neq 0$ (A is called *singular* if $|A| = 0$, *non-singular* otherwise (thus a square matrix has a non-zero determinant iff it is non-singular), and

$$AA^{-1} = A^{-1}A = I :$$

Theorem (Matrix inverse).

inverse = transposed matrix of cofactors over determinant.

Proof. With B as above, $C := AB = (c_{ij})$,

$$c_{ij} = \sum_k a_{ik} b_{kj} = \sum_k a_{ik}.A_{jk}/|A|.$$

If $i = j$, the RHS is 1 (expansion of $|A|$ by its ith row). If not, the RHS is 0 (expansion of the determinant of a matrix with two identical rows). So $c_{ij} = \delta_{ij}$, so $C = AB = I$. Similarly, $BA = I$. //

Solution of linear equations.

If A is $n \times n$, the linear equations

$$Ax = b$$

possess a unique solution x iff A is non-singular (A^{-1} exists), and then

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

If A is singular (A has rank $r < n$), then *either* there is *no* solution (the equations are *inconsistent*), *or* there are *infinitely many* solutions (some equations are *redundant*, and one can give some of the elements x_i *arbitrary* values and solve for the rest – *consistency but non-uniqueness*). What decides between these two cases is the rank of the *augmented* matrix (A, b) obtained by adjoining the vector b as a final column. If $rank(A, b) = rank(A)$, $Ax = b$ is consistent; if $rank(A, b) > rank(A)$, $Ax = b$ is inconsistent.

Orthogonal Matrices.

A square matrix A is *orthogonal* if

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

or equivalently, if

$$A^T A = AA^T = I.$$

Then $|A^T A| = |A^T||A| = |A|.|A| = |I| = 1$, $|A|^2 = 1$, $|A| = \pm 1$ (we take the $+$ sign).

If $A = (a_1, \ldots, a_n)$ (row of column vectors, so A^T is the column of row-vectors a_i^T) is orthogonal, $A^T A = I$, i.e.

$$\begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} (a_1, \ldots, a_n) = I,$$

$a_i^T a_j = \delta_{ij}$: the columns of A are orthogonal to each other, and similarly the rows are orthogonal to each other.

Note. If A, B are orthogonal, so is AB, since $(AB)^T AB = B^T A^T AB = B^T B = I$.

Generalised inverses.

The theory above partially extends to non-square matrices, and matrices not of full rank. For A $m \times n$, call A^- a *generalised inverse* if

$$AA^- A = A.$$

We quote:

1. Generalised inverses always exist (but need not be unique),
2. If the linear equation

$$Ax = b$$

is consistent (has at least one solution), then a particular solution is

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

Eigenvalues and eigenvectors.

If A is square, and

$$Ax = \lambda x \qquad (x \neq 0),$$

λ is called an *eigenvalue* (latent value, characteristic value, e-value) of A, x an *eigenvector* (latent vector, characteristic vector, e-vector) (determined only to within a non-zero scalar factor c, as $A(cx) = \lambda(cx)$). Then

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

has non-zero solutions x, so infinitely many solutions cx, so $A - \lambda I$ is singular:

$$|A - \lambda I| = 0.$$

If A is $n \times n$, this is a polynomial equation of degree n in λ. By the Fundamental Theorem of Algebra (see e.g. M2PM3 L19-L20), there are n roots $\lambda_1, \ldots, \lambda_n$ (possibly complex, counted according to multiplicity).

A matrix A is singular iff the linear equation $Ax = 0$ has some non-zero solution x. This is the condition for 0 to be an eigenvalue:

a matrix is singular iff 0 is an eigenvalue.

Since the coefficient of λ^n in the polynomial $p(\lambda) := |A - \lambda I|$ is $(-)^n$, $p(\lambda)$ factorises as

$$p(\lambda) := |A - \lambda I| = \prod_1^n (\lambda - \lambda_i).$$

Put $\lambda = 0$:

$$|A| = \prod_1^n \lambda_i : \qquad \textit{the determinant is the product of the eigenvalues.}$$

Match the coefficients of $(-\lambda)^{n-1}$: in the RHS, we get a λ_i term for each i, so the coefficient is $\sum_i \lambda_i$, the sum of the eigenvalues. In the LHS, we get an a_{ii} term for each i, so the coefficient is $\sum a_{ii}$, the sum of the diagonal elements of A, which is called the *trace* of A. Comparing:

$$\text{tr } A = \sum_i \lambda_i : \qquad \textit{the trace is the sum of the eigenvalues.}$$

Properties.
1. If A is *symmetric*, eigenvectors x_i, x_j corresponding to distinct eigenvalues λ_i, λ_j are *orthogonal*.
Proof. $Ax_i = \lambda_i x_i$, so $x_i^T A^T = \lambda_i x_i^T$, or $x_i^T A = \lambda_i x_i^T$ as A is symmetric. So $x_i^T A x_j = \lambda_i x_i^T x_j$. Interchanging i and j and transposing (or arguing as above), $x_i^T A x_j = \lambda_j x_i^T x_j$. Subtract: $(\lambda_i - \lambda_j) x_i^T x_j = 0$, so $x_i^T x_j = 0$ as $\lambda_i \neq \lambda_j$. //
2. If A is *real and symmetric*, its eigenvalues are *real*. For $Ax = \lambda x$; taking complex conjugates gives $A\overline{x} = \overline{\lambda}\overline{x}$ as A is real. Transposing, as A is symmetric, this gives $\overline{x}^T A = \overline{\lambda}\overline{x}^T$. So $\overline{x}^T A x = \overline{\lambda}\overline{x}^T x$. Also $Ax = \lambda x$, so $\overline{x}^T A x = \lambda \overline{x}^T x$. Subtract: $0 = (\overline{\lambda} - \lambda)\overline{x}^T x$. But if x has jth element $x_j + iy_j$, $\overline{x}^T x = \sum_j (x_j^2 + y_j^2)$, positive as x is non-zero. So $\overline{\lambda}^T = \lambda$, and λ is real. //
Note. The same proof shows that if A is *anti-symmetric* – $A^T = -A$ – the eigenvalues are purely imaginary.
3. If A is *real and orthogonal*, its eigenvalues are of *unit modulus*: $|\lambda| = 1$.
Proof. If $Ax = \lambda x$, $A\overline{x} = \overline{\lambda}\overline{x}$ as A is real, so $\overline{x}^T A^T = \overline{x}^T \overline{\lambda}$. So $\overline{x}^T A^T A x = \overline{x}^T \overline{\lambda}.\lambda x$, which as A is orthogonal is $\overline{x}^T x = \overline{\lambda}\lambda . \overline{x}^T x$. Divide by $\overline{x}^T x = \sum_i x_i^2 > 0$ (as $x \neq 0$): $\overline{\lambda}.\lambda = |\lambda|^2 = 1$. //
4. If C, A are similar ($C = B^{-1}AB$), A has eigenvalues λ and eigenvectors x – then C has eigenvalues λ and eigenvectors $B^{-1}x$.
Proof. $|A - \lambda I| = 0$, so $|C - \lambda I| = |B^{-1}AB - \lambda B^{-1} I B| = |B^{-1}||A - \lambda I||B| = 0$. So C has eigenvalues λ. $C(B^{-1}x) = (B^{-1}AB)(B^{-1}x) = B^{-1}Ax = B^{-1}\lambda x = \lambda(B^{-1}x)$, so C has eigenvectors $B^{-1}x$. //
Corollary. Similar matrices have the same determinant and trace.
Proof. These are the product and sum of the eigenvalues. //
5. If A is non-singular, the eigenvalues of A^{-1} are the reciprocals λ^{-1} of the eigenvalues λ of A, and the eigenvectors are the same.
Proof. $Ax = \lambda x$, so $x = A^{-1}\lambda x$, so $A^{-1}x = \lambda^{-1}x$. //
6. A is singular iff it has an e-value 0. For, the determinant is the product of the e-values.

Theorem (Spectral Decomposition, or Jordan Decomposition). A symmetric matrix A can be decomposed as

$$A = \Gamma \Lambda \Gamma^T = \sum \lambda_i \gamma_i \gamma_i^T,$$

with $\Lambda = diag(\lambda_i)$ the diagonal matrix of eigenvalues λ_i, $\Gamma = (\gamma_1, \ldots, \gamma_n)$ an orthogonal matrix with columns γ_i standardised eigenvectors ($\gamma_i^T \gamma_i = 1$).

We give a more general result (SVD) below. As a corollary, one can show that for A symmetric, its rank $r(A)$ is the number of non-zero eigenvalues.
Square root of a matrix.

If A is symmetric, with decomposition as above, and we define $\Lambda^{1/2} := diag(\lambda_i^{1/2})$, then putting

$$A^{1/2} := \Gamma\Lambda^{1/2}\Gamma^T,$$

$$\begin{aligned} A^{1/2}A^{1/2} &= \Gamma\Lambda^{1/2}\Gamma^T\Gamma\Lambda^{1/2}\Gamma^T \\ &= \Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma^T \qquad (\Lambda \text{ is orthogonal}) \\ &= \Gamma\Lambda\Gamma^T \qquad (\Lambda = diag(\lambda_i)) \\ &= A. \end{aligned}$$

We call $A^{1/2}$ the *square root* of A. If also A is non-singular (so no eigenvalue is 0, so each λ_i^{-1} is defined), write

$$A^{-1/2} := \Gamma\Lambda^{-1/2}\Gamma^T.$$

A similar argument shows that

$$A^{-1/2}A^{-1/2} = A^{-1},$$

so we call $A^{-1/2}$ the square root of A^{-1}, and the inverse square root of A.
Positive definite matrices.

If A $(n \times n)$ is real and symmetric, A is *positive definite* (respectively *non-negative definite*) if

$$x^TAx > 0 \qquad (\text{respectively } \geq 0) \qquad \text{for all non-zero } x.$$

Here $x^TAx = \sum_{i,j=1}^n x_i a_{ij} x_j = \sum_{i=1}^n a_{ii}x_i^2 + \sum_{i\neq j} a_{ij}x_ix_j$ is a *quadratic form* in the n variables $x_1, \ldots, x_n$ (one can replace $\sum_{i\neq j}$ by $2\sum_{i<j}$).

By the Spectral Decomposition Theorem,

$$\begin{aligned} x^TAx &= x^T\Gamma\Lambda\Gamma^Tx = y^T\Lambda y \qquad (y := \Gamma^Tx) \\ &= \sum \lambda_i y_i^2. \end{aligned}$$

So A is non-negative definite (positive definite) iff $\sum_i \lambda_i y_i^2 \geq 0$ for all y (> 0 for all non-zero y) iff all $\lambda_i \geq 0$ (> 0):

Proposition. A real symmetric matrix A is non-negative definite (positive definite) iff all its eigenvalues are non-negative (positive).

Matrices of the form A^TA are common in Statistics (e.g., in Regression).
1. A^TA is always non-negative definite, since $x^TA^TAx = (Ax)^T(Ax) = y^Ty = \sum y_i^2 \geq 0$, with $y := Ax$. So all eigenvalues of A^TA are non-negative.
2. A^TA is positive definite iff all eigenvalues are positive iff A^TA is non-singular, and one can show this happens iff A has full rank.
3. If $N(A)$ is the null space of A (the vector space of all x with $Ax = 0$), $N(A) = N(A^TA)$.
4. A^TA, A^T and A have the same rank.

2. Singular-values decomposition (SVD).

The following algebraic result is extremely important in Statistics, and in Numerical Analysis. I used [HJ] 3.0, 3.1, [GvL] 2.5; one reference to a standard Linear Algebra book is
S. ROMAN, *Advanced linear algebra*, 3rd ed., Springer, 2008 (or 2nd ed. – not in 1st ed.).
For a statistical treatment, see e.g. Krzanowski [K] (theory, Section 4.1, applications, Ch. 4), or
[R] C. R. RAO, *Linear statistical inference and its applications*, 2nd ed., Wiley, (1973) (1st ed. 1965), 1c(v).
For proof, see there, or SMF 2012 (on course website).

Theorem (Singular-Values Decomposition, SVD). If A $(n \times p)$ has rank r, A can be written

$$A = ULV^T,$$

where U $(n \times r)$ and V $(p \times r)$ are column-orthogonal ($U^TU = V^TV = I_r$) and L $(r \times r)$ is a diagonal matrix with positive elements, and

$$A = \sum_{i=1}^{r} \lambda_i u_i v_i^T,$$

where
(i) the λ_i are the square roots of the positive eigenvalues of A^TA (or AA^T) – the *singular values*;
(ii) the vectors u_i, v_i are eigenvectors of AA^T and A^TA – the *left* and *right singular vectors*.

(For A square and symmetric, this reduces to the Spectral Decomposition).
Eckart-Young Theorem.

The summands $u_i v_i^T$ are of *rank one* (indeed, the general rank-one matrix is of this form). It was shown by C. H. ECKART (1902-73) and G. YOUNG in 1936 that, if the singular values are ranked in order of decreasing size, retaining the first k terms in SVD gives the best approximation (in the sense of a suitable matrix norm – the *Frobenius norm*) to A by a matrix of rank k. The statistical importance of this was studied by I. J. GOOD (1916-2009) in 1969.
Generalised Inverses and SVD.

Recall that the generalised inverse A^- of A satisfies $AA^-A = A$. If A has SVD $A = ULV^T$, one can check that

$$A^- := VL^{-1}U^T$$

is a generalised inverse of A.
Numerical stability.

Part of the practical importance of SVD lies in the fact that it has good numerical stability properties. Small perturbations of a matrix cause only small perturbations of its SVD; thus round-off error etc. has only a limited effect.

3. Statistical setting.

Usually in Statistics we have *univariate data* $x = (x_1, \ldots, x_n)$, and have to analyse it. Sometimes, however, each observation contains several different readings (measurements, for example) on the same 'individual', or object. We then need a two-suffix notation just to describe the data, and so we use matrices throughout.
Notation. We assume that p variables are measured on each of n objects. We assemble the np readings into a *data matrix*

$$X = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix},$$

where x_{ij} is the observation on the jth variable measured on the ith reading.