III: MULTIVARIATE ANALYSIS

1. Preliminaries: Matrix Theory.

Modern Algebra splits into two main parts: Groups, Rings and Fields on the one hand, and Linear Algebra on the other. Linear Algebra deals with *linear transformations* between *vector spaces*. We confine attention here to the *finite-dimensional* case; the infinite-dimensional case needs Functional Analysis and is harder. Broadly, Parametric Statistics can be handled in finitely many dimensions, Non-Parametric Statistics (Ch. VI) needs infinitely many.

Determinants can be traced back to Leibniz (1684, unpublished in his lifetime), Cramer (below) and others; the term first appears in Gauss' thesis *Disquisitiones arithmeticae* in 1801. Although matrices logically precede determinants, they were developed after them. The term is due to J. J. SYLVESTER (1814-1897) in 1850; the theory largely stems from a paper of Arthur CAYLEY (1821-1895) in 1858 (this contains the Cayley-Hamilton Theorem, following work by Hamilton in 1853).

Given a finite-dimensional vector space V, we can always choose a *basis* (a maximal set of linearly independent vectors). All such bases contain the same number of vectors; if this is n, the vector space has *dimension* n.

Given two finite-dimensional vector spaces and a linear transformation α between the two, choice of bases $(e_1, \dots, e_m)$ and $(f_1, \dots, f_n)$ determines a *matrix* $A = (a_{ij})$ by

$$e_i\alpha = \sum_{j=1}^{n} a_{ij} f_j \qquad (i = 1, \dots, m).$$

We write

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix},$$

or $A = (a_{ij})$ more briefly. The a_{ij} are called the *elements* of the matrix; we write A $(m \times n)$ for A (m rows, n columns).

Matrices may be subjected to various operations:

1. *Matrix addition.* If $A = (a_{ij})$, $B = (b_{ij})$ have the same size, then

$$A \pm B := (a_{ij} \pm b_{ij})$$

(this represents $\alpha \pm \beta$ if α, β are the underlying linear transformations).
2. *Scalar multiplication.* If $A = (a_{ij})$ and c is a scalar (real, unless we specify complex), then the matrix

$$cA := (ca_{ij})$$

represents $c\alpha$.
3. *Matrix multiplication.* If A is $m \times n$, B is $n \times p$, then $C := AB$ is $m \times p$, where $C = (c_{ij})$ and

$$c_{ij} := \sum_{k=1}^{n} a_{ik} b_{kj}$$

(this represents the product, or composition, $\alpha\beta$ or $x \mapsto x\alpha\beta$).
Note. Matrix multiplication is non-commutative! – $AB \neq BA$ in general, even when both are defined (which can only happen for A, B square of the same size).
Partitioning.

We may *partition* a matrix A in various ways. for instance, A as above partitions as

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

where A_{11} is $r \times s$, A_{12} is $r \times (n-s)$, A_{21} is $(m-r) \times s$, A_{22} is $(m-r) \times (n-s)$, etc. In the same way, A may be partitioned as
(i) a column of its rows; (ii) a row of its columns.
Rank.

The maximal number of linearly independent *rows* of A is always the same as the maximal number of independent *columns.* This number, r, is called the *rank* of A. When $r = \min(m, n)$ is as big as it could be, the matrix A has *full rank.*
Inverses.

When a square matrix A $(n \times n)$ has full rank n, the linear transformation $\alpha : V \to V$ that it represents is *invertible*, and so has an inverse map $\alpha^{-1} : V \to V$ such that $\alpha\alpha^{-1} = \alpha^{-1}\alpha = i$, the identity map, and α^{-1} is also a linear transformation. The matrix representing α^{-1} is called A^{-1}, the *inverse matrix* of A:

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

the *identity matrix* of size n: $I = (\delta_{ij})$ ($\delta_{ij} = 1$ if $i = j$, 0 otherwise – the *Kronecker delta*).
Transpose.

If $A = (a_{ij})$, the *transpose* is A', or $A^T := (a_{ji})$.

Note that, when all the matrices are defined,

$$(AB)^{-1} = B^{-1}A^{-1}$$

(as this gives $(AB)(AB)^{-1} = ABB^{-1}A^{-1} = AA^{-1} = I$, and similarly $(AB)^{-1}(AB) = I$, as required), and

$$(AB)^T = B^T A^T$$

(the (i, j) element is $\sum_k (B^T)_{ik}(A^T)_{kj} = \sum_k b_{ki}a_{jk} = \sum_k a_{jk}b_{ki} = (AB)_{ji}$).

Determinants.

There are $n!$ *permutations* σ of the set

$$\mathbb{N}_n := \{1, 2, \dots, n\}$$

– bijections $\sigma : \mathbb{N}_n \to \mathbb{N}_n$. Each permutation may be decomposed into a product of *transpositions* (interchanges of two elements), and the *parity* of the number of transpositions in any such decomposition is always the same. Call σ *odd* or *even* according as this number is odd or even. Write

$$\text{sgn } \sigma := 1 \quad \text{if } \sigma \text{ is even}, \quad -1 \quad \text{if } \sigma \text{ is odd}$$

for the *sign* or *signum* of σ. For A a square matrix of size n, the function

$$\det A, \text{ or } |A|, := \sum_\sigma (\text{ sgn } \sigma) a_{1,\sigma(1)} a_{2,\sigma(2)} \dots a_{n,\sigma(n)},$$

where the summation extends over all $n!$ permutations, is called the *determinant* of A, $\det A$ or $|A|$.

Properties.

1. $|A^T| = |A|$.

Proof. If σ^{-1} is the inverse permutation to σ, σ and σ^{-1} have the same parity, so the sums for their determinants have the same terms, in a different order.

2. If two rows (or columns) of A coincide, $|A| = 0$.

Proof. Interchanging two rows changes the sign of $|A|$ (extra transposition, which changes the parity), but leaves A and so $|A|$ unaltered (as the two rows coincide). So $|A| = -|A|$, giving $|A| = 0$.

3. $|A|$ depends linearly on each row (or column) (det is a *multilinear* function, and this area is called Multilinear Algebra).

4. If A is $n \times n$, $|A| = 0$ iff A has rank $r < n$. For then, some row is a linear combination of others. Expanding by this row gives sum of determinants with two rows identical, giving 0.

5. *Multiplication Theorem for Determinants* (Proof: see SMF1415).
If A, B are $n \times n$ (so AB, and BA, are defined),

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

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 it has 0 as an eigenvalue.

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

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

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). Then (C. H. ECKART (1902-73), G. YOUNG in 1936) with the singular values 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, so round-off error etc. is not serious.

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.

As always, n may be large – the larger the better, as large samples are more informative than small ones. The size of p varies with the problem. But typically p might be of the order of 10 or 12, say. A 12-dimensional

'variable space' is unwieldy for many purposes, and we might want a lower-dimensional representation of the data, with as little loss of information as possible. Background: [MKB] Ch. 1, [K] Ch. 1.
Notation.

$$X = (x_{(1)}, \ldots, x_{(p)}) = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

So the *column-vectors* x_i, $x_{(j)}$ relate to the ith *object* and the jth *variable*.
Mean vector. $\overline{x}_i := \frac{1}{n}\sum_{r=1}^n x_{ri}$ is the sample mean of the ith variable; the *sample mean vector* is

$$\overline{x} := \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{pmatrix}.$$

The sample variance s_{ij} between the ith and jth variables is

$$s_{ij} := \frac{1}{n}\sum_{r=1}^n (x_{ri} - \overline{x}_i)(x_{rj} - \overline{x}_j) = \frac{1}{n}\sum_{r=1}^n x_{ri}x_{rj} - \overline{x}_i\overline{x}_j.$$

Form these into a matrix, the *sample covariance matrix* $S := (s_{ij})$:

$$S = \frac{1}{n}\sum_{r=1}^n (x_r - \overline{x})(x_r - \overline{x})^T = \frac{1}{n}\sum_{r=1}^n x_r x_r^T - \overline{x}\ \overline{x}^T.$$

Now $X^T = (x_1, \ldots, x_n)$ (row of columns corresponding to *objects*), so

$$XX^T = (x_1, \ldots, x_n)\begin{pmatrix} x_1^T \\ \vdots \\ x_n \end{pmatrix} = \sum x_r x_r^T.$$

Write $\mathbf{1}$ for a column-vector of n 1s. Then (check) $\mathbf{1}\mathbf{1}^T$ is the $n \times n$ matrix with each element 1, and (check) $X^T\mathbf{1}\mathbf{1}^T X = n^2\overline{x}\ \overline{x}^T$. So

$$S = \frac{1}{n}X^T X - \frac{1}{n^2}X^T\mathbf{1}\mathbf{1}^T X = \frac{1}{n}X^T H X, \quad \text{where} \quad H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

is the $n \times n$ *centring matrix*. We call $M := X^T X = \sum_1^n x_r x_r^T$ the *matrix of sums of squares and products*.

Since s_{ii} is the sample variance of the ith variable, $s_i := \sqrt{s_{ii}}$ is its sample SD. Form the *sample correlation matrix* $R := (r_{ij})$, where

$$r_{ij} := s_{ij}/s_i s_j$$

is the sample correlation coefficient between the ith and jth variables (so $|r_{ij}| \leq 1$). If

$$D := diag(s_i) = diag(\sqrt{s_{ii}}),$$

$$R = D^{-1}SD^{-1}, \qquad S = DRD.$$

One can check:
(i) H is symmetric and idempotent (i.e. $H^2 = H$);
(ii) S is symmetric and non-negative definite;
(iii) R is symmetric and non-negative definite.
Scaling.

If our data is subjected to an affine transformation (change of location and scale) $x \mapsto y := Ax + b$, then (check) $\overline{y} = A\overline{x} + b$, and $S_y = AS_xA^T$. In particular, if

$$y_r := D^{-1}(x_r - \overline{x}) \qquad (*)$$

then Y has mean vector 0 and covariance matrix $D^{-1}S(D^{-1})^T = D^{-1}SD^{-1} = R$, the correlation matrix of X. So the affine transformation $(*)$ *scales* the data X to new data Y, with *zero means* and *unit variances* (1s on the diagonal of S_y – and correlations = covariances r_{ij} of modulus ≤ 1 off the diagonal). This eliminates dependence of the data on arbitrary choices of location and scale in the units, and makes the data *dimensionless*.
Mahalanobis transformation.

Recall that S is non-negative definite, and is positive definite in the typical, or generic, case. Then S^{-1} exists, and hence so do $S^{\pm 1/2}$. If

$$z_r := S^{-1/2}(x_r - \overline{x}) \quad (r = 1, \ldots, n), \qquad (**)$$

then Z has mean vector 0 and covariance matrix $S^{-1/2}SS^{-1/2} = I$. The map $X \mapsto Z$ is the *Mahalonobis transformation*, which not only centres (means to 0) and scales (variances to 1) as above, but also makes the variables *uncorrelated*.
Principal component transformation.

By the Spectral Decomposition Theorem, we can write $S = GLG^T$, where G is an orthogonal matrix and L is a diagonal matrix of eigenvalues of S.

Since S is non-negative definite, its eigenvalues ℓ_i are non-negative, and w.l.o.g. we can re-order the variables so that they decrease in size:

$$\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p \geq 0.$$

The *principal component transformation*

$$y_r := G^T(x_r - \overline{x}) \quad (r = 1, \ldots, n) \qquad (***)$$

takes data X to new data Y, with zero mean and covariance matrix $S_y = G^T S_x G = G^T G L G^T G = L$, as G is orthogonal: $S_y = L$ is diagonal. So the y_r are *uncorrelated* linear combinations of the data, called *principal components*.
R-techniques and Q-techniques.

Multivariate Analysis splits into two broad areas. In the first, we are interested in the p *variables*, that is, in the p *columns* of our data matrix. Methods used here are called *R-techniques*, because they depend on the correlation matrix R. In the second, we are interested in the n *objects*, that is, in the n *rows* of our data matrix. Methods used here are called *Q-techniques*, because they deal directly with the source data (Quelle = source, German).
R-techniques include:

principal components analysis (PCA) [MKB Ch. 8, K 2.3];
factor analysis [MKB Ch. 9, K 16.2];
canonical correlation analysis [MKB Ch. 10, K 14.5].

Q-techniques include:

discriminant analysis [MKB Ch. 11, K 12.3];
cluster analysis [MKB Ch. 13, K 3.1, 9.4];
multidimensional scaling [MKB Ch. 14, K 3.2, 3.3, 9.3].

4. Sample and Population

To describe the population in the p-dimensional case, we need a *population mean (vector)* and a *population covariance (matrix)*:

$$\mu := Ex; \qquad \Sigma := var\ x = E[(x - \mu)(x - \mu)^T].$$

Then (check)

$$E[\overline{x}] = \mu, \qquad var(\overline{x}) = \frac{1}{n}\Sigma, \qquad E[S] = \frac{n-1}{n}.\Sigma.$$

The *unbiased sample covariance matrix* is

$$S_u := \frac{n}{n-1}S;$$

then $E[S_u] = \Sigma$, so S_u is unbiased as an estimator for Σ (as in one dim.).
Objectives.
R-techniques. Here we are interested in the p *variables* (columns of X). If $p = 2$ we can use plots in two dimensions (paper, whiteboard, computer screen); if $p = 3$, we can use our 3-dimensional geometric intuition, and then use computer graphics (based on projective geometry) to represent 3-dimensional reality in 2 dimensions. But if p is 10 or 12, say, it is hard to visualise the data in 10 or 12 dimensions, and so we seek some *lower-dimensional representation* of the data. This will entail some loss of information, which we seek to minimise. We also seek a *parsimonious summarisation* of the data (Principle of Parsimony; Occam's Razor; Einstein's Dictum). One useful technique here is PCA (below). Another is *projection pursuit*.
Q-techniques. Here we are interested in the *objects*. We might want to
(i) represent them as points in space, with closeness corresponding to similarity (multidimensional scaling);
(ii) subdivide or classify into types (cluster analysis);
(iii) assign objects to types (e.g. two types – *discriminant analysis*).
Exploratory Data Analysis (EDA).

As in one dimension, one should begin by 'getting to know the data' by examining it visually. Check for unusual readings (which may be errors – or may be valid and highly informative!), or *outliers*, and decide what to do about any missing readings (e.g. fill in from existing readings – 'imputation').

5. Principal Components Analysis (PCA)

PCA is due to Harold Hotelling (1895-1978) in 1933, following Karl Pearson (1857-1936) in 1901.

We met PCA above in its sample form (see $(***)$); we now turn to the population counterpart of this. We take a random p-vector x, with mean μ and covariance matrix Σ (no distributional assumptions yet). By spectral decomposition of Σ,

$$\Sigma = \Gamma\Lambda\Gamma^T, \qquad \Lambda = \Gamma^T\Sigma\Gamma \qquad (\Sigma = \sum_1^p \lambda_i\gamma_i\gamma_i^T),$$

with $\Lambda = diag(\lambda_i)$, $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ the e-values of Σ, w.l.o.g. in decreasing order, $\Gamma = (\gamma_1, \ldots, \gamma_p)$ the orthogonal matrix of eigenvectors. Write

$$y := \Gamma^T(x - \mu): \qquad y_i = \gamma_i^T(x - \mu),$$

is called the ith *principal component* of x. Then (check)

$$Ey = 0, \qquad var\ y = \Lambda,$$

a diagonal matrix, so the y_i are uncorrelated. Also the $var\ y_i = \lambda_i$ are in decreasing order; their sum and product are the trace and determinant of Σ.

Definition. A linear combination $a^T x = \sum_1^p a_i x_i$ of x is a *standardised linear combination* (SLC) if $\sum_1^p a_i^2 = 1$ (i.e. $a^T a = 1$).

Theorem. The first principal component

$$y_1 = \gamma_1^T (x - \mu)$$

is the SLC of x with the largest variance, λ_1.

Proof. Since $\gamma_i^T \gamma_i = 1$ (the eigenvectors are normalised to have length 1), y_1 is a SLC, and has variance λ_1 by above. If $\alpha := a^T x$ is any other SLC, write

$$a = c_1 \gamma_1 + \ldots + c_p \gamma_p$$

(any p-vector can be written like this, as the columns γ_i are linearly independent, so form a basis). Then

$$var\ \alpha = var(a^T a) = a^T \Sigma a = (\sum_i c_i \gamma_i^T)(\sum_j \lambda_j \gamma_j \gamma_j^T)(\sum_k c_k \gamma_k)$$

$$= \sum_{ijk} c_i \lambda_j c_k \gamma_i^T \gamma_j \gamma_j^T \gamma_k = \sum_{ijk} c_i \lambda_j c_k \delta_{ij} \delta_{jk} = \sum_1^p \lambda_i c_i^2.$$

But $\sum c_i^2 = 1$ and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$, so $var\ \alpha = \sum \lambda_i c_i^2$ is maximised for $c_1 = 1$, $c_i = 0$ for $i = 2, \ldots, p$, when $a = \gamma_1$, and its maximum value is λ_1. //

Note. This choice of $a^T x = \gamma_1^T x$ differs from the first principal component $y_1 = \gamma_1^T (x - \mu)$ only by a constant $\gamma_1^T \mu$, so has the same variance.

Theorem. For each $k = 0, 1, \ldots, p-1$, if $\lambda_k > 0$ the $(k+1)$th principal component

$$y_{k+1} = \gamma_{k+1}^T (x - \mu)$$

is the SLC of x with largest variance uncorrelated with the first k principal components, and this variance is λ_{k+1}.

Proof. If the SLC is $a^T x$ as above, then in the notation above

$$\begin{aligned} cov(a^T x, y_k) &= cov(a^T x, \gamma_k^T (x-\mu)) \\ &= E[(a^T x - E(a^T x)).\gamma_k^T (x-\mu)] \\ &= E[a^T (x-\mu)(x-\mu)^T \gamma_k] \quad (\gamma_k^T (x-\mu) \text{ a scalar, so its own transpose}) \\ &= a^T \Sigma a \qquad (E[(x-\mu)(x-\mu)^T] = \Sigma) \\ &= \sum_1^p c_i \gamma_i \Sigma \gamma_k = \sum_1^p c_i (\Gamma^T \Sigma \Gamma)_{ik}, \end{aligned}$$

which is $\sum c_i \lambda_{ik}$ by spectral decomposition, or $\sum c_i \lambda_i \delta_{ik}$ as Λ is diagonal, which is $c_k \lambda_k$. This is 0 if $a^T x$ is uncorrelated with y_k, but by assumption, $\lambda_k > 0$ (and so $\lambda_1 \geq \ldots \geq \lambda_k > 0$). So $c_k = 0$. Similarly, $c_1 = \ldots = c_{k-1} = 0$. So $a = \sum_{k+1}^p c_i \gamma_i$. As before, $var(a^T x) = \sum_{k+1}^p \lambda_i c_i^2$; as the λ_i are decreasing this is maximised for $c_{k+1} = 1$ and the rest 0, with maximum λ_{k+1}. //

Interpretation. We think of

$$\sum_1^p var\ y_i = \sum_1^p \lambda_i = trace(\Lambda) = trace(\Sigma)$$

as the 'total variability' in the distribution, and $var\ y_1 = \lambda_1$ the 'contribution' of the 1st principal component y_1 to 'explaining' this variability, $var\ y_2 = \lambda_2$ the contribution of y_2, etc. So $\lambda_i/(\lambda_1 + \ldots + \lambda_p)$ is the *proportion* of the total variability explained by the ith principal component, and $(\lambda_1 + \ldots + \lambda_i)/(\lambda_1 + \ldots + \lambda_p)$ is the proportion of the variability explained by the first k PCs. So: if Σ has rank $k < p$, *all* the variability is explained by the first k PCs (the remaining eigenvalues are 0).

How many components to retain?

If we retain k components, there is a trade-off between k large (to explain more variability) and k small (to give a parsimonious representation). We should choose k bearing in mind the *purpose* of our study.

To assist in choice of k, a diagram is often drawn. Plot the points (k, λ_k), or equivalently $(k, \lambda_k/(\sum \lambda_i))$, and join adjacent points by straight-line segments.As the λ_i decrease, the resulting 'broken line' (continuous piecewise-linear function) decreases. We hope to see it decrease steeply at first, then

more slowly, then level off. By analogy with mountain-sides, typically with
(i) the steepest, rocky or cliff, part at the top, then
(ii) a less steep, scree slope in the middle, then
(iii) a gently sloping grassy part below –
such a diagram is called a *scree diagram* (R. B. Cattell (1905-1998) in 1966). Generally we will retain components until somewhere on the scree slope – where depending on how we value parsimony v. accuracy. We may look for an 'elbow', where the gradient flattens out.

Sample principal components

Return to our data matrix X. Let a be a unit p-vector. Then

$$Xa = \begin{pmatrix} x_1^T a \\ \vdots \\ x_n^T a \end{pmatrix}$$

gives n observations of a new variable $x^T a$. The sample variance is $a^T S a$, where S is the sample variance matrix of X; we look for SLCs with maximum variance. Let

$$S = GLG^T$$

be the spectral decomposition of S, $L = diag(l_i)$, with $l_1 \geq \ldots \geq l_p \geq 0$ the e-values of S, $G = (g_1, \ldots, g_p)$ the orthogonal matrix of e-vectors:

$$y_r := G^T(x_r - \overline{x}) \qquad (r = 1, \ldots, n)$$

takes the data matrix X to Y, with mean 0 and covariance matrix L, which is diagonal, so the y_r are *uncorrelated*. Now (check)

$$Y = (X - \mathbf{1}\overline{x}^T)G = (X - \mathbf{1}\overline{x}^T)(g_1, \ldots, g_p), \qquad y_{(k)} = (X - \mathbf{1}\overline{x}^T)g_k$$

gives the SLC of maximal variance, l_k, uncorrelated with $y_{(1)}, \ldots, y_{(k-1)}$. Taking the rth row,

$$y_{rk} = (x_r^T - \overline{x}^T)g_k = g_k^T(x_r - \overline{x}).$$

If the subscript r is unimportant, we can drop it: $y_i = g_i^T(x - \overline{x})$.

Example: Examination scores ([MKB], 1.2.3, Table 1.2.1). This gives data on 88 students' scores on each of 5 Mathematics exams (Mechanics, Vectors, Algebra, Analysis, Statistics); the first two are closed book (C), the last three open book (O). So here $n = 88, p = 5$. The eigenvalues of S are

$$l_1 = 679.2, \quad l_2 = 199.8, \quad l_3 = 102.6, \quad l_4 = 83.7, \quad l_5 = 31.8.$$

The five principal components are found.
1. y_1 gives positive (and comparable) weighting to all 5 marks. This is thus a *weighted average* of the marks, and reflects overall ability (or studiousness – it is difficult to tell these apart from exam performances alone!).
2. y_2 gives positive weight to C and negative weight to O. This is thus a contrast between open-book and closed-book exams. (Students differ greatly, like people generally; most students have a definite preference here; this is often gender-linked).
3. y_3 gives positive weight to Vectors, Algebra and Aalysis, and negative weight to Mechanics and Statistics. This is thus a pure-applied contrast (but would also depend on who taught what!). Again, most students have a definite preference for one or the other.
The last two are less important, as l_4, l_5 are smaller and lack a clear interpretation. We would retain 3 principal components here. We could also use three factors (see above for references to factor analysis).

Similarly for financial stock prices, where the three main factors may be: state of the economy; industrial sector; quality of management.
Covariances v. correlations.

One of the main problems with PCA is that it is *scale-dependent*: the outcome depends on the numbers, hence on the units used. The choice of units is often arbitrary, and then PCA does not have any *intrinsic* meaning. Also PCA looks for SLCs of maximum variability, and the variability can be increased arbitrarily by blowing up the scale in which some variable is measured. So we need to look at and choose the scale of each variable, and this depends on context.

If we use the covariance matrix S, we allow different variables to have differing importance. If we standardise each variance to 1, we pass from S to the correlation matrix R. This is independent of scale and intrinsically meaningful, but now all p variables have the same importance, which may/may not be sensible, depending on context. Moral: think carefully whether to use S or R *before* doing PCA. For more here, see e.g. [K] 2.2.5, esp. p.65-66.