MAST90104: A First Course in Statistical Learning

Notes by Tim Brown, Yao-ban Chan and Owen Jones

Module 2 Linear Algebra

Contents

1 Basics

Linear algebra

Much of the theory of linear models is underpinned by linear algebra (rather unsurprisingly).

We will spend a significant amount of time reviewing some linear algebra results. It might take a while, but it lays the foundation for all of our statistical results.

Much of this material in this section should be familiar - see announcement on LMS sent previously re background and online resources. However, it is always useful to have a reminder, and some things might well be new to you.

References from now on are to Myers and Milton, A First Course in the Theory of Linear Statistical Models (abbreviated MM) and Introduction to Scientific Programming and Simulation Using R (abbreviated spuRs)

1.1 Partitioning - MM 1.1

Partitioning

Matrices can be *partitioned* into smaller (rectangular) *submatrices*:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 3 & -1 \\ 0 & 1 & -1 & 1 \\ 2 & -1 & 0 & 2 \end{bmatrix} = \left[\begin{array}{cc|c|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 3 & -1 \\ \hline 0 & 1 & -1 & 1 \\ 2 & -1 & 0 & 2 \end{array}\right].$$

Partitioning

Partitioned matrices can be manipulated as if the submatrices were single elements (using matrix multiplication instead of scalar multiplication). However, the dimensions of the submatrices must be compatible!

For example, let

$$X = \left[\begin{array}{cc|c} 2 & 1 & 0 \\ \hline 3 & 4 & 1 \end{array}\right] = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right]$$

$$Y = \left[\begin{array}{cc} 1 & 0 \\ 2 & 4 \\ \hline 3 & -1 \end{array}\right] = \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right]$$

Partitioning

Then

$$XY = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right]$$
$$= \left[\begin{array}{c} X_{11}Y_{11} + X_{12}Y_{21} \\ \hline X_{21}Y_{11} + X_{22}Y_{21} \end{array}\right]$$
$$= \left[\begin{array}{c} \left[\begin{array}{cc} 2 & 1 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ 2 & 4 \end{array}\right] + [0] \left[\begin{array}{cc} 3 & -1 \end{array}\right] \\ \hline \left[\begin{array}{cc} 3 & 4 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ 2 & 4 \end{array}\right] + [1] \left[\begin{array}{cc} 3 & -1 \end{array}\right] \end{array}\right]$$

using matrix multiplication for the submatrices.

Partitioning

However, if we partition Y into

$$Y = \left[\begin{array}{cc} 1 & 0 \\ \hline 2 & 4 \\ 3 & -1 \end{array}\right]$$

then we cannot do the multiplication through the partitioning because the components do not have compatible dimensions (for example, X_{11}, Y_{11} are not compatible because X_{11} is 1×2 and Y_{11} is also 1×2) !

1.2 Transposition - MM 1.2

Transposition

The *transpose* of a matrix results when the rows and columns are interchanged. Remember that:

- $(X^T)^T = X$.
- $(XY)^T = Y^T X^T \neq X^T Y^T$!
- A matrix X is *symmetric* if and only if $X^T = X$.

1.3 Inverses - MM 1.3

Identity

The matrix *identity* I is a square matrix of arbitrary size with 1's on the diagonal and 0's off the diagonal:

$$I = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

It has the property that for any X of size $m \times n$,

$$XI_n = I_m X = X,$$

where I_k is the $k \times k$ identity matrix.

Inverse

If X is a square matrix, its *inverse* is the matrix X^{-1} of the same size which satisfies

$$XX^{-1} = X^{-1}X = I.$$

The inverse does not always exist. If it does exist, it is unique and X is called *nonsingular* or *invertible*. Otherwise it is *singular* or *noninvertible* (though some may prefer *vertible*).

X is singular if and only if it has a 0 determinant, $|X| = 0$.

Inverse properties

If X and Y are nonsingular and of the same size, then

- X^{-1} is nonsingular and $(X^{-1})^{-1} = X$.
- XY is nonsingular and $(XY)^{-1} = Y^{-1}X^{-1} \neq X^{-1}Y^{-1}$!
- X^T is nonsingular and $(X^T)^{-1} = (X^{-1})^T$.

1.4 Matrices in R - spuRs 2.4, 2.8

Matrices in R

```
> A <- c(1,2,0,2,3,-1,0,-1,8)
> dim(A) <- c(3,3)
> A
```

```
     [,1] [,2] [,3]
[1,]    1    2    0
[2,]    2    3   -1
[3,]    0   -1    8
```

```
> A <- matrix(c(1,2,0,2,3,-1,0,-1,8),3,3)
> A
```

```
     [,1] [,2] [,3]
[1,]    1    2    0
[2,]    2    3   -1
[3,]    0   -1    8
```

Matrix operations

```
> c <- 2
> c*A
```

```
     [,1] [,2] [,3]
[1,]    2    4    0
[2,]    4    6   -2
[3,]    0   -2   16
```

```
> B <- matrix(c(1,7,-4,8,2,-5,2,2,7),3,3)
> B
```

```
     [,1] [,2] [,3]
[1,]    1    8    2
[2,]    7    2    2
[3,]   -4   -5    7
```

Matrix operations

```
> A+B
```

```
     [,1] [,2] [,3]
[1,]    2   10    2
[2,]    9    5    1
[3,]   -4   -6   15
```

```
> A-B
```

```
     [,1] [,2] [,3]
[1,]    0   -6   -2
[2,]   -5    1   -3
[3,]    4    4    1
```

Matrix operations

```
> A%*%B
```

```
     [,1] [,2] [,3]
[1,]   15   12    6
[2,]   27   27    3
[3,]  -39  -42   54
```

```
> dim(A)
```

```
[1] 3 3
```

```
> det(A)
```

```
[1] -9
```

Matrix operations

```
> A[1,1]
```

```
[1] 1
```

```
> A[c(1,2),c(1,2)]
```

```
     [,1] [,2]
[1,]    1    2
[2,]    2    3
```

```
> A[1,]
```

```
[1] 1 2 0
```

Transposition and identity

```
> t(B)
```

```
     [,1] [,2] [,3]
[1,]    1    7   -4
[2,]    8    2   -5
[3,]    2    2    7
```

```
> I <- diag(3)
> I
```

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Inverse

The R command "solve" solves the matrix equation $A\mathbf{x} = \mathbf{b}$ where A is a square matrix and $\mathbf{x}, \mathbf{b}$ are column vectors. If $\mathbf{b}$ is absent, "solve" finds the inverse of the matrix A.

```
> AI <- solve(A)
> AI

           [,1]       [,2]       [,3]
[1,] -2.5555556  1.7777778  0.2222222
[2,]  1.7777778 -0.8888889 -0.1111111
[3,]  0.2222222 -0.1111111  0.1111111

> AI%*%A

     [,1]         [,2]          [,3]
[1,]    1 2.775558e-17  6.661338e-16
[2,]    0 1.000000e+00 -4.440892e-16
[3,]    0 0.000000e+00  1.000000e+00
```

2 Orthogonality

2.1 What will the work on matrices help us with?

Why?

To calculate confidence intervals, predict future observations and carry out hypothesis tests on linear models, it is necessary to calculate expectations, variances and probabilities for random vectors.

In MAST90105 we did this for the bivariate normal distribution. Recall that linear combinations of bivariate normal random vectors have a normal distribution - and this was the key for many calculations.

Why?

In this subject, the linear models are for an arbitrary number of parameters but linear combinations of *multivariate* random vectors will be at the heart of all the calculations.

Necessarily, these involve matrix operations. The results in this module will all be very helpful in making the calculations of expectations, variances and probabilities possible.

A key aim of this course is to equip you to be able to carry out the calculations on a new model, specially designed for your practical problem in the workplace, with confidence - both on the underlying theory and how to write the code.

2.2 Orthogonal Vectors and Matrices - MM 1.3

Orthogonal vectors

Two $n \times 1$ vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* if and only if

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i = 0.$$

In this case, the geometric depictions of the vectors are perpendicular. (Try this out with vectors in 2 dimensions to recall this if unfamiliar.)

Orthogonal matrices

A square matrix X is *orthogonal* if and only if

$$X^T X = I.$$

If X is orthogonal, then

$$X^{-1} = X^T.$$

Orthogonal matrices

```
> X <- matrix(c(c(1,2,3)/sqrt(14),c(1,1,-1)/sqrt(3),
+               c(5,-4,1)/sqrt(42)),3,3)
> X
          [,1]       [,2]       [,3]
[1,] 0.2672612  0.5773503  0.7715167
[2,] 0.5345225  0.5773503 -0.6172134
[3,] 0.8017837 -0.5773503  0.1543033
> round(t(X)%*%X,5)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Orthogonal matrices

```
> round(X%*%t(X),5)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

2.3 Orthonormal Vectors and Orthogonal Matrices - MM 1.3

Orthogonal and Othonormal

If $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, then $\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2$.

The square root of this quantity, denoted by $\|\mathbf{x}\|$, is called the *norm* or *length* of $\mathbf{x}$.

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is called an *orthonormal set* if and only if each pair of vectors is orthogonal, and each vector has unit length.

Orthogonal vectors

```
> ( x <- c(1,2,3) )

[1] 1 2 3

> ( y <- c(1,1,-1) )

[1]  1  1 -1

> x%*%y

     [,1]
[1,]    0

> t(x)%*%y

     [,1]
[1,]    0
```

Orthogonality

X is an orthogonal matrix if and only if the columns (or rows) of X form an orthonormal set.

```
> X[,1]%*%X[,2]

     [,1]
[1,]    0

> X[1,]%*%X[3,]

              [,1]
[1,] -8.326673e-17
```

3 Eigenthings

3.1 Definition - MM 1.4

Eigenvalues and eigenvectors

Suppose A is a $n \times n$ matrix and $\mathbf{x}$ is a $n \times 1$ nonzero vector which satisfies the equation

$$A\mathbf{x} = \lambda\mathbf{x}$$

where λ is a scalar. Then we say that λ is an *eigenvalue* of A, with associated *eigenvector* $\mathbf{x}$.

Eigenvalues and eigenvectors

Rearranging the definition, we get

$$(A - \lambda I)\mathbf{x} = \mathbf{0}.$$

Now if $A - \lambda I$ is invertible, this produces

$$\mathbf{x} = (A - \lambda I)^{-1}\mathbf{0} = \mathbf{0}.$$

But $\mathbf{x}$ is nonzero by definition, so $A - \lambda I$ must be singular. In particular, its determinant must be 0. Therefore we can find the eigenvalues of a matrix by solving the *characteristic equation* (this is a polynomial in λ)

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

3.2 Example - MM 1.4

Eigenvalue example

Let

$$A = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix}.$$

To find the eigenvalues of A, we solve the equation

$$\begin{vmatrix} 1-\lambda & 1 \\ -2 & 4-\lambda \end{vmatrix} = (1-\lambda)(4-\lambda) - (-2) = 0.$$

This becomes

$$\lambda^2 - 5\lambda + 6 = (\lambda - 2)(\lambda - 3) = 0.$$

Therefore A has two eigenvalues, 2 and 3.

Eigenvalue example

To find the eigenvector(s) of A associated with eigenvalue 2, we solve the system of equations

$$A\mathbf{x} = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{x} = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This is a linear system which has two equations and two unknowns; however, the equations are redundant. Therefore the system has an infinite number of solutions, which always happens for an eigenvector system. One solution is

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Eigenvalue example

```
> A

     [,1] [,2] [,3]
[1,]    1    2    0
[2,]    2    3   -1
[3,]    0   -1    8

> e <- eigen(A)
> e$values

[1]  8.2145852  4.0555651 -0.2701503

> e$vectors

            [,1]      [,2]        [,3]
[1,] -0.05806435 0.5357376  0.84238574
[2,] -0.20945510 0.8184906 -0.53497826
[3,]  0.97609277 0.2075052 -0.06468785
```

Eigenvalue example

```
> det(A - e$values[1]*I)

[1] -2.799516e-14

> A %*% e$vectors[,1]

           [,1]
[1,] -0.4769745
[2,] -1.7205868
[3,]  8.0181972

> e$values[1]*e$vectors[,1]

[1] -0.4769745 -1.7205868  8.0181972
```

3.3 Properties - MM 1.4

Eigenvalue properties

- If A is (real and) symmetric, then its eigenvalues are all real, and its eigenvectors are orthogonal.

- If P is an orthogonal matrix of the same size as A, then the eigenvalues of P^TAP are the same as the eigenvalues of A.

Diagonalization

Theorem 2.1. *Let A be a symmetric $k \times k$ matrix. Then an orthogonal matrix P exists such that*

$$P^TAP = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{bmatrix},$$

where $\lambda_i, i = 1, 2, \dots, k$, are the eigenvalues of A.

Diagonalization

We say that P *diagonalizes* A.

It can be shown that the columns of P must be eigenvectors of A associated with the respective eigenvalues.

Hence (from above) they must be an orthonormal set, which we also know because P is an orthogonal matrix.

Diagonalization example

```
> A

     [,1] [,2] [,3]
[1,]    1    2    0
[2,]    2    3   -1
[3,]    0   -1    8

> e$values

[1]  8.2145852  4.0555651 -0.2701503

> P <- e$vectors
> round(t(P)%*%A%*%P,5)

        [,1]    [,2]     [,3]
[1,] 8.21459 0.00000  0.00000
[2,] 0.00000 4.05557  0.00000
[3,] 0.00000 0.00000 -0.27015
```

4 Rank

4.1 Linear Independence - MM 1.4

Linear independence

Suppose that we have a set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$.

We say that this set is *linearly dependent* if and only if there exists some numbers $a_1, a_2, \ldots, a_k$, which are not all zero, such that

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_k\mathbf{x}_k = \mathbf{0}.$$

If the only way in which this equation is satisfied is for all a's to be zero, then we say that the $\mathbf{x}$'s are *linearly independent.*

Linear independence

If a set of vectors is linearly dependent, then at least one of the vectors can be written as a linear combination of some or all of the rest.

In particular, a set of two vectors is linearly dependent if and only if one of the vectors is a constant multiple of the other.

Linear independence

Example. Are the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

linearly independent?

We substitute them into the equation:

$$\begin{aligned} & a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3 = \mathbf{0} \\ \Rightarrow \quad & \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \end{aligned}$$

so they are linearly independent.

Linear independence

Example. Are the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

linearly independent?

It is easy to see that $\mathbf{x}_3 = \mathbf{x}_1 + \mathbf{x}_2$, that is $\mathbf{x}_1 + \mathbf{x}_2 - \mathbf{x}_3 = \mathbf{0}$. Thus the vectors are linearly dependent.

However, since $\mathbf{x}_1$ is not a constant multiple of $\mathbf{x}_2$, the subset $\{\mathbf{x}_1, \mathbf{x}_2\}$ is linearly independent.

4.2 Definition and Example - MM 1.4

Rank

We define the *rank* of an $n \times k$ matrix by splitting the matrix into columns:

$$X = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_k \end{bmatrix}$$

where $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ are $n \times 1$ vectors. The rank of X, denoted by $r(X)$, is the dimension of the column space of X, i.e. the greatest number of linearly independent vectors in the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$.

It is obvious that $r(X) \leq k$. If $n \geq k$ and $r(X) = k$, we say that X is of *full rank*.

Rank example

```
> (A <- diag(3))

     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1

> library(Matrix)
> rankMatrix(A)[1]

[1] 3
```

Rank example

```
> (B <- matrix(c(2, 3, 1, 1, 0, 2, 3, 3, 3), 3, 3))

     [,1] [,2] [,3]
[1,]    2    1    3
[2,]    3    0    3
[3,]    1    2    3

> rankMatrix(B)[1]

[1] 2
```

4.3 Properties - MM 1.4

Rank properties

- For any matrix X we have $r(X) = r(X^T) = r(X^T X)$.
- If X is $k \times k$, then X is nonsingular if and only if $r(X) = k$.
- If X is $n \times k$, P is $n \times n$ and nonsingular, and Q is $k \times k$ and nonsingular, then $r(X) = r(PX) = r(XQ)$.
- The rank of a diagonal matrix is equal to the number of nonzero diagonal entries in the matrix.
- $r(XY) \leq r(X), r(Y)$.

5 Idempotence

5.1 Definition and Exercise - MM 1.5

Idempotence

We say that a square matrix A is *idempotent* if and only if

$$A^2 = A.$$

Example. The identity matrix I is idempotent.

Exercise. Let X be an $n \times k$ matrix of full rank, $n \geq k$. Show that

$$H = X(X^T X)^{-1} X^T$$

exists and is idempotent.

6 Trace

6.1 Definition MM - 1.5

Trace

The *trace* of a square $k \times k$ matrix X, denoted by $tr(X)$, is the sum of its diagonal entries:

$$tr(X) = \sum_{i=1}^{k} x_{ii}.$$

Example.

$$tr\left(\begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 3 & 2 & -1 \end{bmatrix}\right) = 2 + 1 - 1 = 2.$$

6.2 Properties - MM 1.5

Trace properties

- If c is a scalar, $tr(cX) = c\,tr(X)$.
- $tr(X \pm Y) = tr(X) \pm tr(Y)$.
- If XY and YX both exist, $tr(XY) = tr(YX)$.

Trace properties

Example. Let

$$X = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 3 & 2 \end{bmatrix}, Y = \begin{bmatrix} -1 & 1 & 0 \\ 2 & 4 & 0 \end{bmatrix}.$$

Then

$$tr(XY) = tr\left(\begin{bmatrix} -2 & 2 & 0 \\ 1 & 5 & 0 \\ 1 & 11 & 0 \end{bmatrix}\right) = 3$$
$$tr(YX) = tr\left(\begin{bmatrix} -1 & 1 \\ 8 & 4 \end{bmatrix}\right) = 3$$

so even though $XY \neq YX$, their traces are equal.

7 Theorems

Some linear algebra theorems

7.1 Eigenvalues of idempotent matrices - MM 1.5

Theorem 2.2. *The eigenvalues of idempotent matrices are always either 0 or 1.*

Proof. Let A be an idempotent matrix with eigenvalue λ and associated eigenvector $\mathbf{x}$. By definition,

$$A\mathbf{x} = \lambda\mathbf{x}.$$

Multiplying by A,

$$A^2\mathbf{x} = A\lambda\mathbf{x} = \lambda A\mathbf{x} = \lambda^2\mathbf{x}.$$

But A is idempotent, so

$$\lambda^2\mathbf{x} = A^2\mathbf{x} = A\mathbf{x} = \lambda\mathbf{x}$$
$$(\lambda^2 - \lambda)\mathbf{x} = \mathbf{0}.$$

By definition, $\mathbf{x} \neq \mathbf{0}$, so $\lambda = \lambda^2$. Therefore $\lambda = 0$ or 1.

7.2 Rank and Trace for Symmetric, Idempotent Matrices - MM 1.5

Theorem 2.3. *If A is a symmetric and idempotent matrix, $r(A) = tr(A)$.*

Proof. We take A to be $k \times k$. First we diagonalize A, i.e. find P such that

$$P^T AP = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{bmatrix},$$

where $\lambda_1, \lambda_2, \dots, \lambda_k$ are the eigenvalues of A.
Since P is orthogonal, both P and P^T are nonsingular. Therefore

$$r(P^T AP) = r(P^T A) = r(A).$$

Because $P^T AP$ is diagonal, $r(P^T AP)$ is the number of nonzero eigenvalues of A.

But A is idempotent, so its eigenvalues are either 0 or 1.
To count the number of nonzero eigenvalues, we just need to sum them. But since they are the diagonal elements of $P^T AP$, we can just take its trace.

Therefore

$$r(A) = r(P^T AP) = tr(P^T AP) = tr(PP^T A) = tr(A)$$

since P is orthogonal.

Some linear algebra theorems

7.3 Simultaneous Diagonalisation - MM 1.5

Theorem 2.4. *Let $A_1, A_2, \dots, A_m$ be a collection of symmetric $k \times k$ matrices. Then the following are equivalent:*

- *There exists an orthogonal matrix P such that $P^T A_i P$ is diagonal for all $i = 1, 2, \dots, m$;*
- *$A_i A_j = A_j A_i$ for every pair $i, j = 1, 2, \dots, m$.*

Some linear algebra theorems

Theorem 2.5. *Let $A_1, A_2, \ldots, A_m$ be a collection of symmetric $k \times k$ matrices. Then any two of the following conditions implies the third:*

- *All A_i, $i = 1, 2, \ldots, m$ are idempotent;*
- *$\sum_{i=1}^{m} A_i$ is idempotent;*
- *$A_i A_j = 0$ for $i \neq j$.*

Some linear algebra theorems

Theorem 2.6. *Let $A_1, A_2, \ldots, A_m$ be a collection of symmetric $k \times k$ matrices. If the conditions in Theorem 2.5 are true, then*

$$r\left(\sum_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} r\left(A_i\right).$$

Some linear algebra theorems

Proof.

Consider $\sum_{i=1}^{m} A_i$. By assumption, this matrix is idempotent. As a sum of symmetric matrices it is also symmetric.

Thus by Theorem 2.3,

$$\begin{aligned} r\left(\sum_{i=1}^{m} A_i\right) &= tr\left(\sum_{i=1}^{m} A_i\right) \\ &= \sum_{i=1}^{m} tr(A_i) \\ &= \sum_{i=1}^{m} r(A_i). \end{aligned}$$

7.4 Theorem Examples - MM 1.5

Theorem examples

```
> X <- matrix(c(1/2,1/2,0,1/2,1/2,0,0,0,1),3,3)
> X %*% X

     [,1] [,2] [,3]
[1,]  0.5  0.5    0
[2,]  0.5  0.5    0
[3,]  0.0  0.0    1

> sum(diag(X))

[1] 2
```

Theorem examples

```
> eigen(X)$values
```

```
[1] 1.000000e+00 1.000000e+00 5.551115e-16
```

```
> rankMatrix(X)[1]
```

```
[1] 2
```

Theorem examples

```
> A1 <- matrix(c(1/2,-1/2,-1/2,1/2),2,2)
> A1 %*% A1
```

```
     [,1] [,2]
[1,]  0.5 -0.5
[2,] -0.5  0.5
```

```
> A2 <- matrix(c(1/2,1/2,1/2,1/2),2,2)
> A2 %*% A2
```

```
     [,1] [,2]
[1,]  0.5  0.5
[2,]  0.5  0.5
```

Theorem examples

```
> A1 + A2
```

```
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

```
> (A1 + A2) %*% (A1 + A2)
```

```
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

```
> A1 %*% A2
```

```
     [,1] [,2]
[1,]    0    0
[2,]    0    0
```

Theorem examples

```
> A2 %*% A1
```

```
     [,1] [,2]
[1,]    0    0
[2,]    0    0
```

```
> rankMatrix(A1 + A2)[1]
```

```
[1] 2
```

```
> rankMatrix(A1)[1] + rankMatrix(A2)[1]
```

```
[1] 2
```

8 Quadratic forms

8.1 Definition - MM 2.1

Quadratic forms

Let A be a $k \times k$ matrix and $\mathbf{y}$ a $k \times 1$ vector containing variables.

The quantity

$$q = \mathbf{y}^T A \mathbf{y}$$

is called a *quadratic form* in $\mathbf{y}$, and A is called the matrix of the quadratic form.

Quadratic forms

Note the dimensions: $\mathbf{y}^T$ is $1 \times k$, so q is 1×1. That is, q is a scalar.

In fact it can be expressed as

$$q = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} y_i y_j.$$

Quadratic forms

Example. Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 0 \\ 4 & 6 & 3 \end{bmatrix}.$$

Then

$$\begin{aligned} \mathbf{y}^T A \mathbf{y} &= 2y_1^2 + 3y_1y_2 + y_1y_3 + y_2y_1 + 2y_2^2 + 4y_3y_1 + 6y_3y_2 + 3y_3^2 \\ &= 2y_1^2 + 2y_2^2 + 3y_3^2 + 4y_1y_2 + 5y_1y_3 + 6y_2y_3. \end{aligned}$$

This can be found from either the summation formula or multiplying out the matrices.

8.2 Positive Definiteness - MM 2.1

Positive definiteness

If $\mathbf{y}^T A \mathbf{y} > 0$ for all $\mathbf{y} \neq \mathbf{0}$, then we say that the quadratic form $\mathbf{y}^T A \mathbf{y}$ is *positive definite*; we also say that the matrix A is positive definite.

If $\mathbf{y}^T A \mathbf{y} \geq 0$ for all $\mathbf{y}$, then we say that the quadratic form $\mathbf{y}^T A \mathbf{y}$ is *positive semi-definite*; we also say that the matrix A is positive semi-definite.

Positive definiteness

Example. Let

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

Then

$$\mathbf{y}^T A\mathbf{y} = 2y_1^2 + 2y_2^2 - 2y_1y_2 = y_1^2 + y_2^2 + (y_1 - y_2)^2.$$

The quadratic form will never be negative, and the only way that it can be 0 is if all the squares are 0, i.e. $y_1 = y_2 = 0$. Therefore, $\mathbf{y}^T A\mathbf{y}$ is positive definite.

Positive definiteness theorems

Theorem 2.7. *A symmetric matrix A is positive definite if and only if its eigenvalues are all (strictly) positive.*

Theorem 2.8. *A symmetric matrix A is positive semi-definite if and only if its eigenvalues are all non-negative.*

Positive definiteness

Example. Consider the matrix in the previous example:

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

The eigenvalues of A solve the quadratic equation

$$(2 - \lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1) = 0.$$

Therefore its eigenvalues are 1 and 3, which are both positive, and it is positive definite.

Positive definiteness

```
> A <- matrix(c(2,-1,-1,2),2,2)
> eigen(A)$values

[1] 3 1

> (y <- rnorm(2))

[1] -0.768799191  0.009880679

> t(y)%*%A%*%y

         [,1]
[1,] 1.197492
```

8.3 Differentiation by vectors - MM 2.2

Differentiation of quadratic forms

Occasionally we may wish to differentiate some quadratic forms (e.g. to minimise them). We introduce the idea of differentiation by vectors.

Suppose we have a vector of variables $\mathbf{y} = (y_1, y_2, \ldots, y_k)^T$, and some scalar function of them:

$$z = f(\mathbf{y}).$$

Differentiation of quadratic forms

We define the derivative of z with respect to $\mathbf{y}$ as follows:

$$\frac{\partial z}{\partial \mathbf{y}} = \begin{bmatrix} \partial z/\partial y_1 \\ \partial z/\partial y_2 \\ \vdots \\ \partial z/\partial y_k \end{bmatrix}.$$

Differentiation of quadratic forms

Example. Let

$$A = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 1 & 3 \\ 2 & 3 & 2 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

Consider the quadratic form

$$z = \mathbf{y}^T A \mathbf{y} = y_1^2 + y_2^2 + 2y_3^2 - 2y_1y_2 + 4y_1y_3 + 6y_2y_3.$$

Taking partial derivatives gives

$$\frac{\partial z}{\partial \mathbf{y}} = \begin{bmatrix} 2y_1 - 2y_2 + 4y_3 \\ 2y_2 - 2y_1 + 6y_3 \\ 4y_3 + 4y_1 + 6y_2 \end{bmatrix}.$$

Vector differentiation properties

- If $z = \mathbf{a}^T\mathbf{y}$ where $\mathbf{a}$ is a vector of constants, then $\frac{\partial z}{\partial \mathbf{y}} = \mathbf{a}$.
- If $z = \mathbf{y}^T\mathbf{y}$, then $\frac{\partial z}{\partial \mathbf{y}} = 2\mathbf{y}$.
- If $z = \mathbf{y}^T A\mathbf{y}$, then $\frac{\partial z}{\partial \mathbf{y}} = A\mathbf{y} + A^T\mathbf{y}$. In particular, if A is symmetric, then $\frac{\partial z}{\partial \mathbf{y}} = 2A\mathbf{y}$.

Example. Consider the previous example. We have

$$\begin{aligned} 2A\mathbf{y} &= 2\begin{bmatrix} 1 & -1 & 2 \\ -1 & 1 & 3 \\ 2 & 3 & 2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\ &= 2\begin{bmatrix} y_1 - y_2 + 2y_3 \\ -y_1 + y_2 + 3y_3 \\ 2y_1 + 3y_2 + 2y_3 \end{bmatrix} \\ &= \begin{bmatrix} 2y_1 - 2y_2 + 4y_3 \\ -2y_1 + 2y_2 + 6y_3 \\ 4y_1 + 6y_2 + 4y_3 \end{bmatrix} = \frac{\partial z}{\partial \mathbf{y}}. \end{aligned}$$