Introduction to Matrix Algebra and Multivariate Statistics

Burris

Matrix algebra is incredibly important in statistics, particularly in linear modeling and multivariate analysis. Graduate students matriculate with at least at least a semester of linear algebra experience, but a semester course in linear algebra can vary widely across universities. Some undergraduate linear algebra courses emphasize theory and abstraction to the detriment of working knowledge and application. Others, such as my linear algebra course in undergrad, emphasize computation with concrete examples at the expense of fundamental concepts. Additionally, some students may not have taken linear algebra for several years.

The purpose of this document is to review relevant matrix notation, properties, and concepts in hopes of preparing students for first-year graduate courses in a statistics department. The materials are based off of Haville (2008), Petersen (2012), Gentle (2007), Johnson (2007), and Casella (2008).

1 Matrices: The Basics

1.1 Notation

A real **matrix** (hereafter referred to as a matrix) is a rectangular array of real numbers. The collection of real numbers within a matrix $a_{11}, a_{12}, \ldots, a_{1n}, \ldots, a_{m1}, a_{m2}, \ldots, a_{mn}$ can be arranged as

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \ldots & a_{mn} \end{pmatrix}$$

This matrix, which has m rows and n columns, is an $m \times n$ matrix, where m and n are the **dimensions** of the matrix. The scalar a_{ij} at the intersection of the ith row and jth column of a matrix is called the ijth entry or element. Boldface capital letters (e.g. $\mathbf{A}$) are used to represent matrices.

1.2 Basic Matrix Properties

1.2.1 Addition

The **sum** of two $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} + \mathbf{B}$ and defined to be the $m \times n$ matrix whose ijth element is $a_{ij} + b_{ij}$. For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

Addition is commutative and associative, so

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$$

1.2.2 Multiplication

Let $\mathbf{A}$ be a $m \times n$ matrix and let $\mathbf{B}$ be a $p \times q$ matrix. If $n = p$, the matrix **product AB** is defined to be the $m \times q$ matrix whose ijth element is

$$\mathbf{AB}_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1(5) + 2(7) & 1(6) + 2(8) \\ 3(5) + 4(7) & 3(6) + 4(8) \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

Matrix multiplication is associative, but not commutative in general (**BA** may not even be defined!). That is, if a matrix $\mathbf{C}$ has q rows, then

$$\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$

In addition, matrix multiplication is distributive with respect to addition, so assuming that matrix dimensions are such that all products and sums are defined,

$$\begin{aligned} \mathbf{A}(\mathbf{B} + \mathbf{C}) &= \mathbf{AB} + \mathbf{AC} \\ (\mathbf{A} + \mathbf{B})\mathbf{C} &= \mathbf{AC} + \mathbf{BC} \end{aligned}$$

For any scalar c, we have that

$$\begin{aligned} c(\mathbf{A} + \mathbf{B}) &= c\mathbf{A} + c\mathbf{B} \\ c\mathbf{AB} &= (c\mathbf{A})\mathbf{B} = \mathbf{A}(c\mathbf{B}) \end{aligned}$$

1.2.3 Transposition

The transpose of a $m \times n$ matrix $\mathbf{A}$, denoted by either the symbol $\mathbf{A}'$ or $\mathbf{A}^T$, is the $n \times m$ matrix whose ijth element is the jith element of $\mathbf{A}$. For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

Below are some properties of matrix transposes:

$$(\mathbf{A}^T)^T = \mathbf{A}$$

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

1.3 Square Matrices

A matrix that has the same number of rows as columns is called a **square** matrix. A $n \times n$ square matrix is said to be of order n. The n elements $a_{11}, a_{22}, \ldots, a_{nn}$ of

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

that fall on the imaginary diagonal line extending from the top left corner to the bottom right corner of the matrix are known as **diagonal elements**. Any elements of the matrix that are not on the main diagonal are called **off-diagonal elements**. Note that the product $\mathbf{AA}$ is only defined if A is square. If it is, we can write $\mathbf{A}^2 = \mathbf{AA}$ and $\mathbf{A}^k = \mathbf{AAA} \cdots \mathbf{A}$.

1.3.1 Symmetric Matrices

A square matrix $\mathbf{A}$ is symmetric if $\mathbf{A}^T = \mathbf{A}$. For example, the matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{pmatrix}$$

is symmetric.

1.3.2 Diagonal Matrices

A square matrix $\mathbf{A}$ is a **diagonal matrix** if all of its off-diagonal elements are equal to 0. That is,

$$\mathbf{A} = \begin{pmatrix} d_1 & 0 & \ldots & 0 \\ 0 & d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & d_n \end{pmatrix}$$

for some $d_1, d_2, \ldots, d_n$. Diagonal matrices are very easy to work with, since for a $m \times n$ matrix $\mathbf{A}$ and $n \times n$ diagonal matrix $\mathbf{D}$, the ijth element of $\mathbf{DA}$ (or $\mathbf{AD}$) is equal to $d_i a_{ij}$.

1.3.3 Identity Matrix

The diagonal matrix

$$diag(1, 1, \ldots, 1) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

with diagonal elements all equal to 1 is the identity matrix. We use the symbol $\mathbf{I}_n$ (or just $\mathbf{I}$ if the context is clear) to denote the identity matrix. For any matrix $\mathbf{A}$,

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

1.3.4 Triangular Matrices

If all the elements of a square matrix that are located below and to the left of the diagonal are 0, then the matrix is called **upper triangular**. An upper triangular matrix looks like this:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

Similarly, a matrix with all elements above and to the right of the main diagonal equal to zero is a **lower triangular matrix**. More formally, an $n \times n$ matrix $\mathbf{A}$ is upper triangular if $a_{ij} = 0$ for $j < i = 1, 2, \dots, n$. Triangular matrices are very easy to work with, by hand and computationally, and much of the matrix decompositions that we'll learn about later involve decomposing a matrix into the product of one or more triangular matrices.

1.3.5 Trace of a Square Matrix

The **trace** of a square matrix $\mathbf{A}$ of order n is defined to be the sum of the n diagonal elements of $\mathbf{A}$. This is denoted by the symbol $tr(\mathbf{A})$. Thus,

$$tr(\mathbf{A}) = a_{11} + a_{22} + \ldots + a_{nn}$$

Some properties of the trace are

$$tr(k\mathbf{A}) = k \ tr(\mathbf{A})$$

$$tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$$

$$tr(\mathbf{A}^T) = tr(\mathbf{A})$$

$$tr(\mathbf{AB}) = tr(\mathbf{BA})$$

$$tr(\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_k) = tr(\mathbf{A}_k\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_{k-1})$$

1.4 Vectors

A matrix with only one column, that is, a matrix of the form

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$$

is a **column vector**. Similarly, a matrix with one row is called a **row vector**. We distinguish vectors from matrices by referring to them by a boldface lowercase letter. So **a** is a column vector, and $\mathbf{a}^T$ is a row vector. The symbol $\mathbf{1}_m$ denotes a column vector with all entries equal to one.

The **inner product** of two vectors $\mathbf{a}$ and $\mathbf{b}$ is $\mathbf{a}^T\mathbf{b}$. The sum of a vector $\mathbf{a}$ can be written as $\mathbf{1}^T\mathbf{a}$. The mean of a vector is

$$\mathbf{1}^T\mathbf{a}/n = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{a}$$

If $\mathbf{a}^T\mathbf{b} = 0$, then $\mathbf{a}$ and $\mathbf{b}$ are **perpendicular** to one another. A collection of vectors is **orthogonal** if and only if they are pairwise perpendicular.

A vector $\mathbf{a}$ is a **unit vector** if $\mathbf{a}^T\mathbf{a} = 1$.

Null vectors (and null matrices) are denoted by the notation $\mathbf{0}_m$.

1.5 Exercises

1. Show, using the definition of a matrix product, that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.
2. Show that the product of two upper-triangular matrices is also upper triangular. What is $\mathbf{AB}_{ii}$?
3. Give an example of two symmetric matrices whose product is not symmetric. When will the product of two symmetric matrices be symmetric?

2 Matrices: Beyond Basic Operations

Over this next section, we'll go into more detail about matrices. It is here that we'll review foundational concepts such as inverses, determinants, matrix calculus, and eigendecompositions.

2.1 Partitioned Matrices

A **submatrix** of a matrix $\mathbf{A}$ can be obtained by striking out rows and/or columns of $\mathbf{A}$. A **partitioned matrix** is a $m \times n$ matrix $\mathbf{A}$ that has been expressed in the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1c} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \dots & \mathbf{A}_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{r1} & \mathbf{A}_{r2} & \dots & \mathbf{A}_{rc} \end{pmatrix}$$

where the submatrix $\mathbf{A}_{ij}$ is referred to as the ijth block of $\mathbf{A}$. Partitioning matrices can be very useful, particularly when dealing with large sparse matrices (matrices that have many entries equal to zero).

2.2 Linear Independence

A nonempty finite set $\{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A_k}\}$ of $m \times n$ matrices is **linearly dependent** if there exist scalars $x_1, x_2, \dots, x_k$, not all equal to zero, such that

$$x_1\mathbf{A}_1 + x_2\mathbf{A_2} + \cdots + x_k\mathbf{A}_k = \mathbf{0}$$

If no such scalars exist, the set is said to be **linearly independent**. You may have only seen the notion of linear dependence only applied to vectors, but this definition is a little more general, since a column vector is just a matrix with one column.

In addition, a set of two or more $m \times n$ matrices is linearly dependent if and only if at least one of the matrices can be expressed as a linear combination of the other matrices ($\mathbf{A_j}$ is expressible as a linear combination of $\mathbf{A_1}, \ldots, \mathbf{A}_{j-1}, \mathbf{A}_{j+1}, \ldots \mathbf{A_k}$).

2.3 Subspaces, Basis, and Rank

2.3.1 Linear Spaces

The **column space** of a $m \times n$ $\mathbf{A}$ is the set whose elements consist of all m dimensional vectors that can be expressed as linear combinations of the n columns of $\mathbf{A}$. The elements of the column space of $\mathbf{A}$ are all m dimensional vectors of the form

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n$$

where $x_1, x_2, \ldots, x_n$ are scalars and $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are the columns of $\mathbf{A}$. For example, the column space of the 3×4 matrix

$$\begin{pmatrix} 2 & -4 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

includes the column vector

$$\begin{pmatrix} 4 \\ -2 \\ -3 \end{pmatrix} = 2\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} + 0\begin{pmatrix} -4 \\ 2 \\ 0 \end{pmatrix} - 3\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + 0\begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}$$

but not the column vector $(2, 0, 0)^T$. The row space is defined similarly, by finding all n dimensional vectors that can be expressed as a linear combination of the rows of $\mathbf{A}$.

Row spaces and column spaces of a matrix are examples of a more general concept called a **linear space**. A nonempty set $\mathcal{V}$ is a linear space if it is closed under element-wise multiplication and scalar multiplication (i.e. $\mathbf{A}$, $\mathbf{B} \in \mathcal{V} \Rightarrow \mathbf{A} + \mathbf{B} \in \mathcal{V}, k\mathbf{A} \in \mathcal{V}$).

A subset $\mathcal{U}$ of a linear space $\mathcal{V}$ is called a subspace of $\mathcal{V}$ if $\mathcal{U}$ is a linear space. For example, the column space $\mathcal{C}(\mathbf{A})$ of a $m \times n$ matrix $\mathbf{A}$ is a subspace of $\mathbb{R}^m$ and the row space $\mathcal{C}(\mathbf{A})$ of $\mathbf{A}$ is a subspace of $\mathbb{R}^n$.

2.3.2 Bases

The **span** of a finite, nonempty set of matrices $S = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A_k}\}$ is the set of all matrices that are expressible as a linear combination of the elements of S. The span is denoted by $\text{sp}(S)$ and is a linear space.

A **basis** for a linear space $\mathcal{V}$ is a finite set of linearly independent matrices in $\mathcal{V}$ that spans $\mathcal{V}$. Note that if the columns of $\mathbf{A}$ are linearly independent, then the columns of $\mathbf{A}$ are a basis for $\mathcal{C}(\mathbf{A})$.

For example, $\{(1,0,0)^T, (0,1,0)^T, (0,0,1)^T\}$ and $\{(1,1,0)^T, (1,2,0)^T, (0,2,3)^T\}$ are both bases for $\mathbb{R}^3$. However, $\{(1,0,0)^T, (0,1,0)^T, (0,0,1)^T, (0,2,3)^T\}$ is not a basis for $\mathbb{R}^3$ since the set is linearly dependent.

2.3.3 Rank

The number of matrices in a basis for a linear space $\mathcal{V}$ is called the **dimension** of $\mathcal{V}$ and is denoted by $\dim(\mathcal{V})$. The **row rank** of a matrix $\mathbf{A}$ is the dimension of the row space of $\mathbf{A}$. Similarly, the **column rank** of a matrix $\mathbf{A}$ is the dimension of the column space of $\mathbf{A}$.

An important result in matrix algebra is the row rank and the column of a matrix are equal. This common value is called the **rank** of $\mathbf{A}$ and denoted by $\text{rank}(\mathbf{A})$.

A matrix is said to have **full row rank** if $\text{rank}(\mathbf{A}) = m$ and **full column rank** if $\text{rank}(\mathbf{A}) = n$. Square matrices ($m = n$) are said to be **full rank** or **nonsingular** if it has both full row rank and full column rank. A $n \times n$ matrix with rank less than n is **singular**. For example,

$$\mathbf{A} = \begin{pmatrix} 1 & 3 & 4 \\ 2 & 6 & 8 \\ 5 & 7 & 12 \end{pmatrix}$$

is a singular matrix, since $\text{rank}(\mathbf{A}) = 2$, but $\mathbf{A}$ is a 3×3 matrix.

2.4 Matrix Inverses

2.4.1 Definition

A $n \times n$ matrix $\mathbf{A}$ is invertible if and only if $\mathbf{A}$ is nonsingular. If $\mathbf{A}$ is invertible, then there exists a unique inverse matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{\mathbf{-1}} = \mathbf{A}^{\mathbf{-1}}\mathbf{A} = \mathbf{I}$.

For any $n \times n$ nonsingular diagonal matrix $\mathbf{D}$,

$$\mathbf{D}^{-1} = diag(1/d_1, 1/d_2, \ldots, 1/d_n)$$

The inverse of a 2×2 nonsingular matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$\mathbf{A}^{-1} = (1/k) \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

where $k = a_{22}a_{11} - a_{12}a_{21}$.

2.4.2 Properties

Let $\mathbf{A}$ be a nonsingular $n \times n$ matrix. Then

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$$

$$(k\mathbf{A})^{-1} = 1/k\mathbf{A}^{-1}$$

for any non-zero scalar k. If $\mathbf{B}$ is also a nonsingular $n \times n$ matrix, then

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

and more generally,

$$(\mathbf{A_1}\mathbf{A_2} \cdots \mathbf{A_k})^{-1} = \mathbf{A_k}^{-1} \cdots \mathbf{A_2}^{-1}\mathbf{A}_1^{-1}$$

2.4.3 Orthogonal Matrices

A nonsingular matrix $\mathbf{A}$ is an **orthogonal matrix** if $\mathbf{A}^T = \mathbf{A}^{-1}$. Equivalently, $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$

Examples of orthogonal matrices include the identity matrix $\mathbf{I}_n$ and, for any angle θ, the 2×2 matrix

$$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$$

2.4.4 Generalized Inverses

A **generalized inverse** or **pseudoinverse** of an $m \times n$ matrix $\mathbf{A}$ is any $n \times m$ matrix $\mathbf{G}$ such that

$$\mathbf{AGA} = \mathbf{A}$$

.

If $\mathbf{A}$ is nonsingular, then it has a unique generalized inverse. Otherwise, $\mathbf{A}$ has an infinite number of generalized inverses. There are many properties of generalized inverses that are useful when solving a linear system, but they are not presented here for brevity.

2.5 Determinants

The **determinant** of a square $n \times n$ matrix $\mathbf{A}$ is denoted by either $|\mathbf{A}|$ or $\det(\mathbf{A})$. The minor $\mathbf{M}_{ij}$ of element A_{ij} is the $n-1 \times n-1$ matrix that is formed by removing the ith row and jth column from $\mathbf{A}$. The cofactor of A_{ij} is $C_{ij} = (-1)^{i+j}|\mathbf{M}_{ij}$. Expanding along the ith row,

$$|\mathbf{A}| = \sum_{j=1}^{n} A_{ij}C_{ij}$$

Note that $|\mathbf{A}^T| = |\mathbf{A}|$ and that $|k\mathbf{A}| = k^n|\mathbf{A}|$. $\mathbf{A}$ is singular if $|\mathbf{A}| = 0$ and nonsingular otherwise.

2.6 Projection Matrices

A square matrix $\mathbf{A}$ is **idempotent** if $\mathbf{A}^2 = \mathbf{A}$. A square matrix $\mathbf{P}$ is a **projection matrix** if and only if $\mathbf{P}$ is idempotent. If $\mathbf{P}$ is also orthogonal, then $\mathbf{P}$ is called an orthogonal projector.

A very useful property of projection matrices is that their rank is equal to their trace, namely

$$tr(\mathbf{A}) = rank(\mathbf{A})$$

The identity matrix $\mathbf{I}_n$ is the only full rank idempotent matrix.

A primary application of projection matrices is linear models and the method of least squares. One important result is that the projection $\mathbf{z}$ of a n-dimensional vector $\mathbf{y}$ onto the column space of a $n \times p$ matrix $\mathbf{X}$ is

$$\mathbf{z} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T y}$$

Letting $\hat{\beta}$ be the ordinary least squares estimate of β, where $\mathbf{y} = \mathbf{X}\beta + \epsilon$, we have that the fitted values $\mathbf{X}\hat{\beta}$ is the projection of the response vector $\mathbf{y}$ onto the column space of the covariate matrix $\mathbf{X}$.

2.7 Quadratic Forms

If $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ is an n-dimensional vector, then a **quadratic form** is

$$\mathbf{x}^T\mathbf{A}\mathbf{x}$$

$\mathbf{A}$ is **positive definite** if, for all $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$. If instead $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$, then $\mathbf{A}$ is **nonnegative definite**.

Symmetric nonnegative definite matrices are encountered frequently in linear models and other areas of statistics. In particular, the covariance matrix (we'll get to that next section) is nonnegative definite. In addition, any positive definite matrix is nonsingular.

2.8 Eigenvalues and Eigenvectors

A scalar λ is said to be an **eigenvalue** of an $n \times n$ matrix $\mathbf{A}$ if there exists a non-null vector $\mathbf{x}$ such that

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

The set of eigenvalues is called the **spectrum** of $\mathbf{A}$. All eigenvalues of positive-definite matrices are positive and all eigenvalues of nonnegative-definite matrices are nonnegative.

A non-null vector $\mathbf{x}$ is an **eigenvector** of $\mathbf{A}$ if there exists a scalar λ such that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.

If $\mathbf{A}$ is a symmetric $n \times n$ matrix, eigenvectors $\mathbf{v}_j$ and $\mathbf{v}_k$ associated with distinct eigenvalues $\lambda_j \neq \lambda_k$ are orthogonal. In addition, if $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v_n})$, then

$$\mathbf{AV} = \mathbf{V}\boldsymbol{\Lambda}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix with the n eigenvalues on the diagonal and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. The spectral theorem expresses $\mathbf{A}$ as a weighted average of rank 1 matrices,

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T = \sum_{j=1}^{n} \lambda_j \mathbf{v}_j \mathbf{v}_j^T$$

This decomposition of a square matrix into its eigenvalues and eigenvectors is called the **eigendecomposition**. This decomposition will always exist if the matrix is symmetric. The rank of $\mathbf{A}$ is the number of nonzero eigenvalues, the trace of $\mathbf{A}$ is

$$tr(\mathbf{A}) = \sum_{j=1}^{n} \lambda_j$$

and the determinant is

$$|\mathbf{A}| = \prod_{j=1}^{n} \lambda_j$$

2.9 Matrix Decompositions

Decomposing matrices into products of easy to work with matrices is very useful when implementing matrix algorithms and examining properties of multivariate data. We've already seen one example with the eigendecomposition, which is particularly relevant for a multivariate statistical technique called principal components analysis (PCA). While the ones presented here are not exhaustive, they're particularly relevant in statistical applications.

2.9.1 Cholesky Decomposition

The Cholesky decomposition takes a positive definite matrix $\mathbf{A}$ and decomposes it into $\mathbf{U}^T\mathbf{U}$, where $\mathbf{U}$ is an upper triangular matrix. Cholesky factorization is substantially faster than the more general LU factorization because pivoting is not required.

2.9.2 QR Decomposition

The QR decomposition decomposes a general $m \times n$ matrix $\mathbf{A}$ into $\mathbf{QR}$, where $\mathbf{Q}$ is an orthogonal $m \times m$ matrix and $\mathbf{R}$ is an upper triangular $m \times n$ matrix. This decomposition is useful for solving least squares problems and numerically approximating eigenvectors.

2.9.3 Singular Value Decomposition

A $m \times n$ matrix $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \mathbf{UDV}^T$$

where $\mathbf{U}$ is an $m \times n$ matrix such that $\mathbf{U}^T\mathbf{U} = \mathbf{I_m}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I_n}$, and $\mathbf{D}$ is a nonnegative diagonal matrix with entries $d_1 \geq \cdots \geq d_n$.

This representation is called the **singular value decomposition** of $\mathbf{A}$ (or SVD). The SVD provides a description of the within-row variation via $\mathbf{V}$ and a description of the within column variation via $\mathbf{U}$. This is very useful when describing large data matrices with a low-dimensional approximation.

The SVD of a symmetric matrix is identical to its eigendecomposition.

2.10 Matrix Calculus

Let $\mathbf{x} = (x_1, \ldots, x_K)^T$ be a k dimensional vector and let

$$\mathbf{y} = (y_1, \ldots, y_J)^T = (f_1(\mathbf{x}), \ldots, f_J(\mathbf{x}))^T = \mathbf{f}(\mathbf{x})$$

be a J dimensional vector, where $\mathbf{f} : \mathbb{R}^K \rightarrow \mathbb{R}^J$. Then the partial derivative of $\mathbf{y}$ with respect to $\mathbf{x}^T$ yields the $J \times K$ **Jacobian matrix**

$$\mathbf{J_x y} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}^T} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_K} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_J}{\partial x_1} & \frac{\partial y_J}{\partial x_2} & \cdots & \frac{\partial y_J}{\partial x_K} \end{pmatrix}$$

The **Jacobian** of the tranformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ is

$$J = |\mathbf{J_x y}|$$

If $y = f(\mathbf{x})$ is a scalar, then the **gradient** vector is

$$\Delta_x y = \frac{\partial y}{\partial \mathbf{x}} = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots, \frac{\partial y}{\partial x_K}\right)^T = \left(\frac{\partial y}{\partial \mathbf{x}^T}\right)^T = (\mathbf{J}_x y)^T$$

The $K \times K$ matrix of second order partial derivatives is called the **Hessian matrix**. This is defined mathematically as

$$\mathbf{H_x} y = \frac{\partial}{\partial \mathbf{x}}\left(\frac{\partial y}{\partial \mathbf{x}}\right)^T = \frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^T} = \begin{pmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1 x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 x_K} \\ \frac{\partial^2 y}{\partial x_1 x_2} & \frac{\partial^2 y}{\partial x_2^2} & \cdots & \frac{\partial^2 y}{\partial x_2 x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_1 x_K} & \frac{\partial^2 y}{\partial x_2 x_K} & \cdots & \frac{\partial^2 y}{\partial x_K^2} \end{pmatrix}$$

The gradient and Hessian, if they can be found in closed form, are very useful when optimizing a scalar function $y = f(\mathbf{x})$.

The derivative of a $J \times K$ matrix $\mathbf{A}$ with respect to an r dimensional vector x is the $(Jr \times K)$ matrix of derivatives of $\mathbf{A}$ with respect to each element of $\mathbf{x}$. In other words,

$$\frac{\partial \mathbf{A}}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{A}^T}{\partial x_1}, \ldots, \frac{\partial \mathbf{A}^T}{\partial x_r}\right)^T$$

Similar to normal differentiation, we have the following rules:

$$\begin{aligned}
\frac{\partial(\alpha \mathbf{A})}{\partial \mathbf{x}} &= \alpha \frac{\partial \mathbf{A}}{\partial \mathbf{x}} \\
\frac{\partial(\mathbf{A}+\mathbf{B})}{\partial \mathbf{x}} &= \frac{\partial \mathbf{A}}{\partial \mathbf{x}} + \frac{\partial \mathbf{B}}{\partial \mathbf{x}} \\
\frac{\partial(\mathbf{A}\mathbf{B})}{\partial \mathbf{x}} &= \left(\frac{\partial \mathbf{A}}{\partial \mathbf{x}}\right)\mathbf{B} + \mathbf{A}\left(\frac{\partial \mathbf{B}}{\partial \mathbf{x}}\right)
\end{aligned}$$

2.11 Exercises

1. Show that the product of two orthogonal matrices is also orthogonal, assuming that the product is nonsingular.

2. Show that if $\mathbf{P}$ is a projection matrix, then $\mathbf{I} - \mathbf{P}$ is also a projection matrix.

3. Show that the eigenvalues of an idempotent matrix are either 0 or 1.

4. Suppose the $\mathbf{Y}$ is an $n \times p$ data matrix. There exists a centering matrix $\mathbf{C}$ such that $\mathbf{CY}$ subtracts the column means of $\mathbf{Y}$ from $\mathbf{Y}$. As such, the column means of $\mathbf{CY}$ are all zero. Find $\mathbf{C}$. Is $\mathbf{C}$ a projection matrix?

5. What is the Jacobian matrix for the polar coordinate transformation? ($x = r\cos(\theta)$, $y = r\sin(\theta)$)

3 Random Matrices: Multivariate Statistics

In the morning session, we mainly looked at univariate and bivariate random variables. However, often we will have many observations and many variables for each observations in a data set. In this section, we will review some multivariate probability theory, with particular emphasis on the multivariate normal distribution.

3.1 Random Vectors

If we have d random variables $X_1, X_2, \ldots, X_d$, each defined on the real line, we can write them as the d dimensional column vector

$$\mathbf{X} = (X_1, \cdots X_d)^T$$

which we call a d-dimensional **random vector**. The joint distribution function of the random vector $\mathbf{X}$ is

$$\begin{aligned}
F_X(\mathbf{x}) &= F_X(x_1, \ldots, x_d) \\
&= P(X_1 \le x_1, \ldots, X_d \le x_d) \\
&= P(\mathbf{X} \le \mathbf{x})
\end{aligned}$$

If F_X is absolutely continuous, then the joint density function f_X of $\mathbf{X}$ is

$$f_X(\mathbf{x}) = f_X(x_1, \ldots, x_d) = \frac{\partial^d F_X(x_1, \ldots, x_d)}{\partial x_1 \cdots \partial x_d}$$

You can recover the cumulative distribution function from the density function by integrating

$$F_X(\mathbf{x}) = \int_{-\infty}^{x_d} \cdots \int_{-\infty}^{x_1} f_X(u_1, \ldots, u_d) du_1 \cdots du_d$$

To find the marginal density of a subset of the d variables, you can just integrate the others out. For example, if we have a joint bivariate density $f_{X_1,X_2}(x_1, x_2)$, then

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) dx_2 \qquad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) dx_1$$

The components of a random vector $\mathbf{X}$ are **independent** if the joint distribution function is a product of the marginal distribution functions

$$F_X(\mathbf{x}) = \prod_{i=1}^{d} F_i(x_i)$$

In addition, under independence, the joint density is a product of the marginal densities

$$f_X(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i)$$

3.2 Multivariate Moments

If $\mathbf{X}$ is a random vector with values in $\mathbb{R}^d$, then its expected value is given by the d dimensional vector

$$\mu_X = E(\mathbf{X}) = (E(X_1), \cdots, E(X_d)) = (\mu_1, \cdots, \mu_d)^T$$

and the $d \times d$ **covariance matrix** of $\mathbf{X}$ is

$$\begin{aligned}
\mathbf{\Sigma}_{XX} &= \text{cov}(\mathbf{X}, \mathbf{X}) \\
&= E[(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)^T] \\
&= E[(X_1 - \mu_1, \cdots, X_d - \mu_d)(X_1 - \mu_1, \cdots, X_d - \mu_d)^T] \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \ldots & \sigma_d^2 \end{pmatrix}
\end{aligned}$$

where

$$\sigma_i^2 = \text{var}(X_i) = E[(X_i - \mu_i)^2]$$

is the variance of X_i and

$$\sigma_{ij} = \text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

is the covariance between X_i and X_j.

The **correlation matrix** of $\mathbf{X}$ can be obtained by from $\mathbf{\Sigma}_{XX}$ by dividing the ith row by σ_i and the jth column by σ_j. The $d \times d$ matrix is then

$$P_{XX} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1d} \\ \rho_{21} & 1 & \dots & \rho_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d1} & \rho_{d2} & \dots & 1 \end{pmatrix}$$

where

$$\rho_{ij} = \rho_{ji} = \begin{cases} \frac{\sigma_{ij}}{\sigma_i \sigma_j} & i \neq j \\ 1 & \text{otherwise} \end{cases}$$

is the pairwise correlation coefficient between X_i and X_j. The correlation coefficient will always lie between -1 and 1 and is a measure of association between X_i and X_j.

If $\mathbf{Y}$ is a linear function of $\mathbf{X}$ such that

$$\mathbf{Y} = \mathbf{AX} + \mathbf{b}$$

the mean vector and covariance matrix of $\mathbf{Y}$ is given by

$$\mu_Y = \mathbf{A}\mu_x + \mathbf{b}$$
$$\mathbf{\Sigma}_{YY} = \mathbf{A}\mathbf{\Sigma}_{\mathbf{XX}}\mathbf{A}^T$$

3.3 Multivariate Normal Distribution

The multivariate normal distribution is a generalization of the normal distribution to two or more dimensions. Although data rarely completely follows a multivariate normal distribution, it can provide a reasonable approximation to reality, particularly if **mixtures** of multivariate normal distributions are used.

Recall that the density of the univariate normal distribution is given by

$$f(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The form of the multivariate normal looks similar, but slightly different. A random d vector $\mathbf{X}$ follows a multivariate normal distribution with mean vector μ and positive definite symmetric covariance matrix $\mathbf{\Sigma}$ if it has the density function

$$f(\mathbf{x}|\mu, \mathbf{\Sigma}) = (2\pi)^{-d/2} |\mathbf{\Sigma}|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

We notationally denote a d dimensional normal distribution as

$$\mathbf{X} \sim N_d(\mu, \mathbf{\Sigma})$$

The **Mahalanobis distance** from $\mathbf{x}$ to μ is given by the quadratic form

$$\Delta = \sqrt{(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

An important result is that a random vector $\mathbf{X}$ follows a multivariate distribution if and only if every linear function of $\mathbf{X}$ follows a univariate normal distribution.

In linear models, we often assume that $\mathbf{\Sigma} = \sigma^2 \mathbf{I_d}$, in which case the density function reduces to

$$f(\mathbf{x}|\mu, \sigma) = (2\pi\sigma)^{-d/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)}$$

3.3.1 Conditional Normal Distribution

Suppose we have two random vectors $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ has d_1 components and $\mathbf{Y}$ had d_2 components. Let $\mathbf{Z}$ be the random $d_1 + d_2$ vector

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$$

Then the expected value of $\mathbf{Z}$ is given by

$$\mu_Z = E[\mathbf{Z}] = \begin{pmatrix} E[\mathbf{X}] \\ E[\mathbf{Y}] \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

and the covariance matrix of $\mathbf{Z}$ is the partitioned $(d_1 + d_2 \times d_1 + d_2)$ matrix

$$\begin{aligned} \mathbf{\Sigma}_{ZZ} &= \begin{pmatrix} cov(\mathbf{X},\mathbf{X}) & cov(\mathbf{X},\mathbf{Y}) \\ cov(\mathbf{Y},\mathbf{X}) & cov(\mathbf{Y},\mathbf{Y}) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{\Sigma_{XX}} & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_{YY} \end{pmatrix} \end{aligned}$$

where $\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_{YX}^T$. The marginal distribution of $\mathbf{Y}$ is

$$\mathbf{Y} \sim N_{d_2}(\mu_y, \mathbf{\Sigma}_{YY}$$

The conditional distribution of $\mathbf{Y}$ given that $\mathbf{X} = \mathbf{x}$ is multivariate normal with mean vector and covariance matrix given by

$$\begin{aligned} \mu_{Y|X} &= \mu_Y + \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}(\mathbf{x} - \mu_X) \\ \mathbf{\Sigma}_{Y|X} &= \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY} \end{aligned}$$

3.4 Wishart Distribution

Given n independent and identically distributed d vectors

$$\mathbf{X}_i \sim N_d(\mu, \mathbf{\Sigma})$$

we say that the random positive-definite, symmetric matrix

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}^T$$

follows a **Wishart distribution** with n degrees of freedom and matrix $\mathbf{\Sigma}$. We denote the Wishart distribution by

$$\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{\Sigma})$$

The Wishart distribution is important in Bayesian statistics because it can serve as a conjugate prior distribution for a covariance matrix if the data is assumed to come from a multivariate normal distribution. You can think of the Wishart as a randomly drawn covariance matrix multiplied by the degrees of freedom n, since $E[\mathbf{W}] = n\mathbf{\Sigma}$. As $n \to \infty$, $\mathbf{W}/n \to \mathbf{\Sigma}$.

There are several useful properties of the Wishart distribution. I'll list a couple of the main ones below.

1. Let $\mathbf{W}_j \sim \mathcal{W}_d(n_j, \mathbf{\Sigma})$ be independent. Then $\sum_{j=1}^{m} \mathbf{W}_j \sim \mathcal{W}_d(\sum_{j=1}^{m} n_j, \mathbf{\Sigma})$
2. Suppose $\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{\Sigma})$ and let $\mathbf{A}$ be a constant matrix having full row rank. Then $\mathbf{AWA}^T \sim \mathcal{W}_d(n, \mathbf{A\Sigma A}^T)$.
3. Suppose $\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{\Sigma})$ and let $\mathbf{a}$ be a fixed d dimensional vector. Then $\mathbf{a}^T\mathbf{Wa} \sim (\mathbf{a}^T\mathbf{\Sigma a})\chi_n^2$.

You can think of the Wishart as a multidimensional chi-square distribution. If $\mathbf{W}$ follows a Wishart distribution, then $\mathbf{W}^{-1}$ follows an **inverse Wishart distribution**.

3.5 Exercises

Given that

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3(\mu, \mathbf{\Sigma})$$

where

$$\mu = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

1. Find the correlation matrix ρ of $\mathbf{X}$
2. Find the marginal distribution of X_2.
3. Find the marginal distribution of $\{X_1, X_3\}$.
4. Find the conditional distribution of $X_1 | X_3 = -1$.
5. Find the conditional distribution of $X_1 | \{X_2 = 1, X_3 = -1\}$
6. Are $\{X_1, X_3\}$ and X_2 independent?
7. Are $X_1 + X_2$ and $X_1 - X_2$ independent?
8. Let $\mathbf{Y} = \mathbf{AX} + \mathbf{a}$ where $\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix}$ and $\mathbf{a} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Find the distribution of $\mathbf{Y}$.
9. What is the distribution of $W = (\mathbf{X} - \mu)(\mathbf{X} - \mu)^T$?

4 References

Much of the material and exercises in this handout comes from the following books. I highly recommend consulting these books throughout your time in graduate school. Being familiar with matrices is essential when trying to understand statistical literature.

- David Harville. *Matrix Algebra from a Statistician's Perspective*
- James Gentle. *Matrix Algebra: Theory, Computations, and Applications in Statistics*
- Kaare Petersen and Michael Pedersen. *The Matrix Cookbook*
- Richard Johnson and Dean Wichern. *Applied Multivariate Statistical Analysis*
- Alan Izenman. *Modern Multivariate Statistical Techniques*