BIOS 835: Basic Matrix Algebra, Random Vectors, and the Covariance Matrix

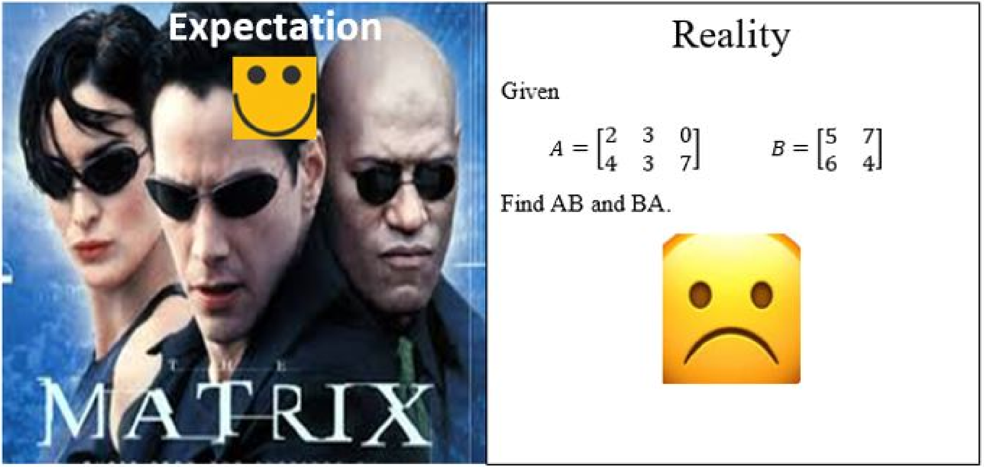
Alexander McLain

August 29, 2023

Outline

1. Basic Matrix Algebra
2. Random Vectors
3. Covariance Matrices

What is the Matrix?



Why linear algebra?

- Linear algebra is a branch of mathematics concerning:
  - linear equations,
  - linear functions, and
  - and their representations in vector spaces and through matrices
- Linear algebra is fundamental to many algorithms and concepts in machine learning; for example,
  - **Data Representation:** Datasets in machine learning are often represented as matrices.
  - **Transformations and Dimensionality Reduction:** Techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) use linear algebra to project data into lower-dimensional spaces.

Why linear algebra?

- **Eigenvalues and Eigenvectors:** These are used in clustering, PCA, and in many optimization algorithms.
- **Distance and Similarity Computations:** Computing the distance between vectors or matrices is fundamental for many areas of ML.
- **Linear Regression:** Linear algebra is fundamental.
- **Regularization:** Techniques like L1 (Lasso) and L2 (Ridge) regularization in regression analysis can be described and implemented using linear algebra.
- **Support Vector Machines (SVMs):** utilize dot products, particularly with the kernel trick.
- **Neural Networks and Deep Learning:** Neural network operations involve matrix multiplications, activations, and transformations.

Introduction

- Some helpful resources:
  - Check out the **Matrix Algebra Tutorial** here,
  - A free text *"Linear Algebra"* is available here.
- Let $\boldsymbol{A}$ be a $m \times n$ matrix.
- Let a_{ij} be the element in row i column j

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

  here $\boldsymbol{A}$ is a 2×3 matrix.
- Some special cases:
  - a Vector: $m \times 1$ or $1 \times n$:
  - a square matrix when $m = n$ (we'll deal with these often).
  - an identity matrix $\boldsymbol{I}$.
- Transpose of $\boldsymbol{A}$, $\boldsymbol{A}'$.

Basic Matrix Operation

- Matrix addition and subtraction: $\boldsymbol{A} + \boldsymbol{B}$ and $\boldsymbol{A} - \boldsymbol{B}$
- General properties:
  - $a\boldsymbol{A} = \boldsymbol{A}a$ where a is a scalar.
  - $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$
  - $a(\boldsymbol{A} + \boldsymbol{B}) = a\boldsymbol{A} + a\boldsymbol{B}$
  - $a(\boldsymbol{A} + \boldsymbol{B})' = a\boldsymbol{A}' + a\boldsymbol{B}'$
- Matrix multiplication, $\boldsymbol{AB}$
- General properties:
  - $\boldsymbol{C}(\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{CA} + \boldsymbol{CB}$
  - $\boldsymbol{IA} = \boldsymbol{A}$
  - $\boldsymbol{A}'\boldsymbol{A}$ is a square matrix
  - $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$

Linear Dependence

- The columns $c_1, \ldots, c_k$ of a matrix are linearly dependent if there exists a set of scalar values $\lambda_1, \ldots, \lambda_k$ (at least one is none zero) such that

$$\lambda_1 c_1 + \ldots + \lambda_k c_k = 0$$

- An example

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 1 & 5 \\ 2 & 3 & 1 \end{bmatrix}$$

- Linearly independent: if the only set of λ_j values to satisfy the above equation is a set of all zeros.

Rank

- The **rank** of a matrix is a fundamental concept in linear algebra, and it holds significant importance in various machine-learning contexts.
- **Definition:**
  - The rank of a matrix A is the maximum number of linearly independent row vectors (or equivalently, column vectors) in the matrix. It provides a measure of the "information content" of the matrix.
- Mathematically, if a matrix has a rank r, it means that:
  1. There are r linearly independent rows (or columns) in the matrix.
  2. Any other row (or column) can be represented as a linear combination of these r rows (or columns).
- If a matrix X is $n \times p$ with $p \gg n$, what is the maximum rank of X.

Determinant and Inverse Matrix

- The determinant of $\boldsymbol{A}$ is denoted by $det(\boldsymbol{A}) = |\boldsymbol{A}|$.
- The inverse of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^{-1}$.
- Example, let

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

then

  - $|\boldsymbol{A}| = ad - bc$,
  - and

$$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Determinant and Inverse Matrix

Some properties of the determinants and inverses:

- $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ This is the main property of the inverse (check for the above $\boldsymbol{A}$).
- $|\boldsymbol{A}| = |\boldsymbol{A}'|$, and $|\boldsymbol{A}| = 1/|\boldsymbol{A}^{-1}|$
- $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$
- $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
- $(\boldsymbol{A}')^{-1} = (\boldsymbol{A}^{-1})'$

Some required properties for $\boldsymbol{A}$ to have an inverse are: $\boldsymbol{A}$ is linearly independent, square matrix, non-zero determinant, and full rank.

Positive Definite Matrix

- A matrix A is said to be symmetric positive definite if it meets the following criteria:
  1. **Symmetry:** The matrix A is symmetric, which means it is equal to its transpose, $A = A^T$
  2. **Positive Definiteness:** For any non-zero column vector x, the scalar $x^T Ax$ is positive.
- Some of the useful properties of these include:
  - All eigenvalues of a symmetric positive definite matrix are positive.
  - It is invertible, and its inverse is also symmetric positive definite.

Orthogonal and Orthonormal Matrices

- A square matrix $\boldsymbol{Q}$ is called orthogonal if its transpose is its inverse: $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = I$ where I is the identity matrix of the same order as $\boldsymbol{Q}$.
- Properties of orthogonal matrices:
  - The columns (and rows) of an orthogonal matrix form an orthogonal set of vectors, meaning their dot product is zero
  - The determinant of an orthogonal matrix is either 1 or -1.
  - The inverse of an orthogonal matrix is also orthogonal.
- An orthonormal matrix is an orthogonal matrix, but one where each vector has a unit length (norm of 1).

Matrix Decompositions

LR Decomposition

- $\boldsymbol{A} = \boldsymbol{LR}$ with $\boldsymbol{L}-$lower-triangular and $\boldsymbol{R}-$upper-triangular

Cholesky Decomposition

- If $\boldsymbol{A}$ is symmetric positive definite $\boldsymbol{A} = \boldsymbol{LL}'$

QR Decomposition

- $\boldsymbol{A} = \boldsymbol{QR}$ where $\boldsymbol{Q}$ is an orthogonal matrix, i.e., $det(\boldsymbol{Q}) = 1$

Eigenvalues and Eigenvectors

- $\boldsymbol{A}$ is $J \times J$.
- A scalar λ is called an eigenvalue of $\boldsymbol{A}$ then there is a nontrivial solution $\boldsymbol{x}$ to $\boldsymbol{Ax} = \lambda\boldsymbol{x}$. Such an $\boldsymbol{x}$ is called an eigenvector corresponding to the eigenvalue λ.
  - Note that if $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ then $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$
- If $\boldsymbol{A}$ is positive definite then $\boldsymbol{A}$ will have J eigenvalues where $\lambda_j > 0$ for all $j = 1, \ldots, J$.
  - and are usually ordered such that $\lambda_1 > \lambda_2 > \ldots > \lambda_J$
- Let $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ be eigenvectors from $\lambda_j \neq \lambda_k$ then
  - $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ are orthogonal, i.e. $\boldsymbol{x}_j'\boldsymbol{x}_k = \boldsymbol{0}$

Spectral theorm

- For a given real symmetric matrix A, there exists $\boldsymbol{AQ} = \boldsymbol{Q}\Lambda$ or

$$\boldsymbol{A} = \boldsymbol{Q}\Lambda\boldsymbol{Q}' = \sum_{j=1}^{J} \lambda_j \boldsymbol{q}_j \boldsymbol{q}_j',$$

  where
  - $\boldsymbol{Q}$ is a $J \times J$ matrix with $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_J$, i.e., $\boldsymbol{Q}$ is orthogonal, and
  - $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_J)$ and $\lambda_1 > \lambda_2 > \ldots > \lambda_J$.
- $trace(\boldsymbol{A}) = \sum_j \lambda_j$
- $det(\boldsymbol{A}) = \prod_j \lambda_j$

Singular Value Decomposition

- Singular Value Decomposition (SVD) of a matrix A of size $m \times n$ is given by:

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

Where:

- U (of size $m \times m$) is the left singular vector matrix. Its columns are the eigenvectors of AA^T.
- Σ (of size $m \times n$) is a diagonal matrix. The elements on its diagonal are the singular values of A, and they are non-negative and usually sorted in descending order. These values are the square roots of the eigenvalues of $A^T A$ (or equivalently, AA^T).
- V^T (of size $n \times n$) is the right singular vector matrix. Its rows are the eigenvectors of $A^T A$.

Functions of matrices

- $\boldsymbol{A}$ is $J \times J$ and $\phi : \mathbb{R}^J \Rightarrow \mathbb{R}^J$ then

$$\phi(\boldsymbol{A}) = \sum_{j=1}^{J} \phi(\lambda_j)\boldsymbol{x}_j\boldsymbol{x}_j'$$

  where λ_j and $\boldsymbol{x}_j$ are normalized to have norm 1.
- Example, $\boldsymbol{A}^{1/2} = \sum_{j=1}^{J} \sqrt{\lambda_j}\boldsymbol{x}_j\boldsymbol{x}_j'$ or $\boldsymbol{A}^{-1} = \sum_{j=1}^{J} \frac{1}{\lambda_j}\boldsymbol{x}_j\boldsymbol{x}_j'$

Matrix Norms

- $\boldsymbol{A}$ is $J \times J$, how do we measure the size of $\boldsymbol{A}$?

Properties of a Norm $||\cdot||$

1. $||\boldsymbol{A}|| \geq 0$
2. $||\boldsymbol{A}|| = 0 \Leftrightarrow \boldsymbol{A} = \boldsymbol{0}$
3. $||\boldsymbol{A} + \boldsymbol{B}|| \leq ||\boldsymbol{A}|| + ||\boldsymbol{B}||$
4. $||\alpha\boldsymbol{A}|| = |\alpha|||\boldsymbol{A}||$

Examples of Matrix Norms

- Frobenius Norm: often referred to as the Euclidean norm for matrices, it is the square root of the sum of the absolute squares of its elements.

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

  where A is an $m \times n$ matrix.
- Lp Norm: for vectors (a special case of matrices), the Lp norm is defined as:

$$||v||_p = \left(\sum_i |v_i|^p \right)^{\frac{1}{p}}$$

- Max Norm: the maximum absolute row sum of the matrix. For vectors, it's the maximum absolute value of the elements.

Examples of Matrix Norms

- Spectral Norm: for a matrix A with singular value decomposition $A = U\Sigma V^T$, the spectral norm is the largest entry in Σ.
- Nuclear (or trace) Norm: the sum of the singular values of a matrix.
- Condition Number: (not actually a norm) is the ratio of its largest singular value to its smallest singular value.
  - gives an indication of the numerical stability of matrix inversion and the sensitivity of the system's solution to changes in the input.

Matrix Calculus

- $\boldsymbol{y}$: J-vector
- $\boldsymbol{x}$: K-vector
- Let $f(\boldsymbol{x}) = \boldsymbol{y}$ be a mapping (i.e., a function) from $\mathbb{R}^K \to \mathbb{R}^J$.
- The partial of $\boldsymbol{y}$ with respect to $\boldsymbol{x}$ is

$$J_x\boldsymbol{y} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{dy_1}{dx_1} & \cdots & \frac{dy_1}{dx_K} \\ \vdots & \ddots & \vdots \\ \frac{dy_J}{dx_1} & \cdots & \frac{dy_J}{dx_K} \end{pmatrix}_{J\times K}$$

which is call the Jacobian matrix

Matrix Calculus

- $\boldsymbol{y}$: 1-vector
- $\boldsymbol{x}$: K-vector
- Hessian Matrix of $\boldsymbol{y}$ with respect to $\boldsymbol{x}$ is

$$H_x y = \frac{\partial^2 \boldsymbol{y}}{\partial \boldsymbol{x}^2} = \begin{pmatrix} \frac{dy}{dx_1^2} & \cdots & \frac{dy_1}{dx_1 dx_K} \\ \vdots & \ddots & \vdots \\ \frac{dy}{dx_1 dx_K} & \cdots & \frac{dy}{dx_K^2} \end{pmatrix}_{K \times K}$$

Taylor series approximation

- $\boldsymbol{y}$: 1-vector
- $\boldsymbol{x}$: K-vector
- Let $f(\boldsymbol{x}) = \boldsymbol{y}$ be a mapping (i.e., a function) from $\mathbb{R}^K \to \mathbb{R}$.
- The Taylor series approximation of $f(\boldsymbol{x})$ at $\boldsymbol{c}$ is

$$f(\boldsymbol{x}) = f(\boldsymbol{c}) + [J_x f(\boldsymbol{c})](\boldsymbol{x} - \boldsymbol{c}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{c})^T [H_x f(\boldsymbol{c})](\boldsymbol{x} - \boldsymbol{c})$$

Random Vectors

Example: 10-week-old guinea pigs

- $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)$
  - X_1 - weight
  - X_2 - length
  - X_3 - cholesterol
  - X_4 - time on maze test
  - X_5 - Wheel distance

Multivariate Distributions

- $F_x(\boldsymbol{x}) = F_x(x_1, x_2, x_3, x_4, x_5) = \Pr(X_1 \leq x_1, X_2 \leq x_2, X_3 \leq x_3, X_4 \leq x_4, X_5 \leq x_5)$
- If $\boldsymbol{X}$ is all continuous

$$f_x(\boldsymbol{x}) = \frac{\partial^5 F_x(\boldsymbol{x})}{\partial x_1 \partial x_2 \cdots \partial x_5}$$

where

$$F_x(\boldsymbol{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_5} f_x(\boldsymbol{x}) dx_1 dx_2 \cdots dx_5$$

- If $\boldsymbol{X}$ is all discrete

$$p_x(\boldsymbol{x}) = \Pr(X_1 = x_1, X_2 = x_2, \cdots X_5 = x_5)$$

Expectation and Variance

- In general $E(\boldsymbol{X}) = \boldsymbol{\mu}$.
- For each element of $\boldsymbol{X}$ we have

$$E(X_j) = \mu_j, \quad \text{var}(X_j) = E\{(X_j - \mu_j)^2\} = \sigma_j^2$$

- If we think of the elements of $\boldsymbol{X}$ together, we must consider the possible relationship among different elements of $\boldsymbol{X}$. This leads us to covariance.

Covariance Matix

- Covariance: a measure of how two random variables vary together.
- Mathematically we have,

$$\mathrm{cov}(X_j, X_k) = E\{(X_j - \mu_j)(X_k - \mu_k)\}$$

- The covariance matrix of a random vector is defined by $E\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})\}$

$$= \begin{pmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \dots & E(X_1 - \mu_1)(X_n - \mu_n) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 & \dots & E(X_2 - \mu_2)(X_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_n - \mu_n)(X_1 - \mu_1) & E(X_n - \mu_n)(X_2 - \mu_2) & \dots & E(X_n - \mu_n)^2 \end{pmatrix}$$

Covariance Matix

- Let $\text{cov}(X_j, X_k) = E(X_j - \mu_j)(X_k - \mu_k) = \sigma_{jk}$ with $\sigma_{jj} = \sigma_j^2$

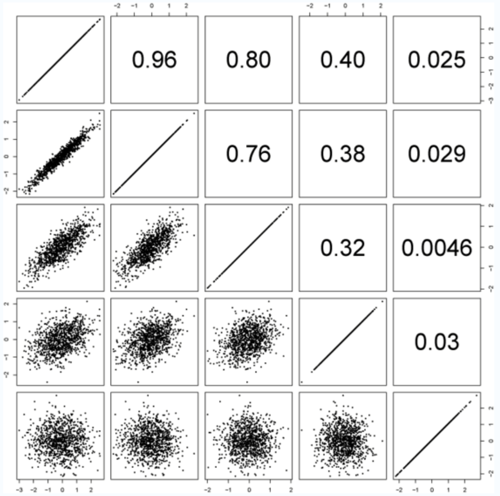
$$\boldsymbol{\Sigma} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})\} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix}$$

- The population correlation of two elements is

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_j^2 \sigma_k^2}}$$

with corresponding correlation matrix.

Correlation Matix Example



Multivariate Normal Distribution

- If a random vector $\boldsymbol{X}$ has a multivariate normal distribution we write this as $\boldsymbol{X} \sim MVN_n(\boldsymbol{\mu}, \Sigma)$ or $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ is the mean and Σ is the covariance.
- The density can be written as

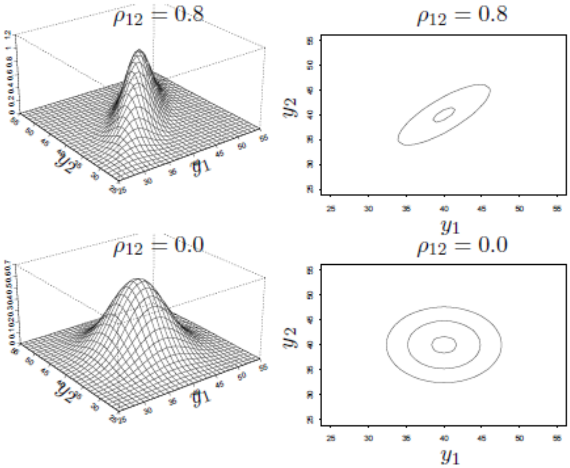
$$f(\boldsymbol{X}) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{-(\boldsymbol{X}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})/2\right\}$$

- A simple case is the bivariate normal (where n=2)

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

and Σ^{-1} can be found using the equation for the inverse given above.

Multivariate Normal Distribution



Random Matricies

- Let $\boldsymbol{X} \sim MVN_r(\mathbf{0}, \Sigma)$
- We have data $\boldsymbol{X}_i$ for $i = 1, \ldots, n$ where $\boldsymbol{X}_i$ is an r-vector.
- Let

$$W = n\hat{\Sigma} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T$$

- Then W has a central Wishart distribution with n degrees of freedom and associated matrix Σ.
- Written as $W \sim W_r(n, \Sigma)$.