LA - PROJECT

Linear Algebra and its application in Genetics

ARYAN GARG, ABHIRAM TILAK & SHREYAS ADIGA

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The fusion of genetics and linear algebra has led to remarkable advancements in understanding and manipulating biological systems. Drawing inspiration from the pioneering work of Gregor Mendel and his model of inheritance, this project aims to delve into the profound applications of linear algebra in unraveling the complexities of genetic phenomena. By utilizing concepts such as matrices, vectors,eigenvalues,eigenvectors and diagonalization , we can mathematically model intricate biological processes, including prey-predator dynamics, gene expression, evolution, and population genetics.

INTRODUCTION

This Paper briefly covers various concepts related to genetics, and tries to discuss the computational applications for the following models :-

- Mathematical aspects of Genetics
- Population Genetics and evolution using Markov Chains and Lotka-Volterra equations

- Studying the Mendel's postulates and Gene-expression
- Gene-interaction and Disease Susceptibility

STATE OF ART LITERATURE

The landscape of genetics research is continuously evolving, with promising future directions on the horizon. The following potential innovations hold significant promise for further advancements in the field: While your project focuses on Logistic regression on CpG chains, PCA, and disease susceptibility, there are several potential future innovations within these areas that hold promise for further advancements:

- **Machine Learning and Deep Learning:** Leveraging the power of machine learning algorithms, such as neural networks, convolutional neural networks (CNNs) [12], and recurrent neural networks (RNNs) [11], can enhance our ability to extract meaningful insights from vast genetic datasets, providing more accurate disease predictions and therapeutic targets.
- **Advanced Feature Selection Techniques:** Enhancing the feature selection process within Logistic regression on CpG chains can lead to improved model performance and interpretability. (read section 6 for more) Future innovations may involve exploring novel feature selection methods specifically tailored for genetic data, such as recursive feature elimination (RFE), lasso regularization, or genetic algorithm-based approaches.
- **Integration of DNA Methylation Data (DMeth):** DNA methylation, an essential epigenetic modification, plays a crucial role in gene regulation. Future research could involve incorporating DMeth data into Logistic regression models on CpG chains, allowing for a more comprehensive analysis of the interplay between DNA methylation patterns and gene expression.
- **Dimensionality Reduction Techniques for Genetic Data: [8]** In addition to PCA, there are other dimensionality reduction methods worth exploring. For instance, non-linear dimensionality reduction techniques like t-SNE (t-Distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold Approximation and Projection) can capture complex genetic structures, potentially revealing subtle relationships and clusters within the data.

1 GENE EXPRESSION

Gene expression refers to the process by which the information encoded in a gene is used to synthesize a functional gene product, such as a protein or RNA molecule. Gene expression is a fundamental process in living organisms and is tightly regulated to ensure proper development, growth, and functioning of cells and tissues.

1.1 Transcription

It can be divided into 2 processes:

- Transcription : Transcription is the process by which the information in a strand of DNA is copied into a new molecule of messenger RNA (mRNA). Transcription occurs in the nucleus of eukaryotic cells and the cytoplasm of prokaryotic cells. It begins with the process of initiation , in which an enzyme called RNA polymerase binds to a specific region on the DNA called the promoter. The promoter provides a signal for the start of transcription and helps position the RNA polymerase at the appropriate location on the DNA strand. Then it proceeds to elongation phase , wherein RNA polymerase unwinds a part of the DNA , and moves along the DNA double strand and adds complimentary RNA nucleotides(Adenine-Uracil and Guanine-Cytosine). Transcription ends with Termination , which occurs when RNA polymerase reaches a termination sequence on the DNA. At this point, the RNA polymerase and the newly synthesized RNA molecule are released from the DNA template.
- Translation : Translation refers to the process by which the genetic information stored in mRNA (messenger RNA) molecules is used to synthesize proteins. It occurs in the ribosomes, which are cellular structures responsible for protein synthesis.

1.2 Mendelian Genetics

Mendelian Genetics refers to the basic principles of inheritance first described by Gregor Mendel. Mendel's work laid the foundation for our understanding of how traits are passed from one generation to the next.

Mendel formulated three fundamental laws based on his experiments with pea plants.

- Law of Segregation : Every individual possesses two alleles for a given trait, and these alleles segregate (separate) during gamete formation, with each gamete receiving only one allele.
- Law of Independent Assortment : The alleles for different traits segregate independently of one another during gamete formation, leading to various combinations of traits in offspring.
- Law of Dominance : In a heterozygous individual (having two different alleles), the dominant allele is expressed in the phenotype , while the recessive allele remains hidden.

We denote genotypes of a particular trait , say Tallness of pea plants, with alleles T and t with combinations of them.TT denotes a plant with both the alleles having Tall allele (homozygous dominant); Tt denotes a plant with 1 tall allele and 1 short allele (heterozygous); tt denotes a plant with 2 short alleles (homozygous recessive) .

Consider this table which denotes the fraction of all offsprings if one parent's genotype is homozygous dominant and the other parent is anything.

Genotype of Offspring	Genotype of Parents		
	RR-RR	RR-Rr	RR-rr
RR	1	$\frac{1}{2}$	0
Rr	0	$\frac{1}{2}$	1
rr	0	0	0

Table 1: Genotypes of Offsprings in Mendelian Pea Plants

We will represent this information as a 3×3 stochastic matrix as follows. [9]

$$M = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Let a_n ,b_n , c_n denote fraction of TT,Tt and tt plants in n^{th} generation . Thus , we have the following equations, representing the n^{th} generation.

$$a_n = a_{n-1} + \frac{1}{2} \cdot b_{n-1}$$
$$b_n = c_{n-1} + \frac{1}{2} \cdot b_{n-1}$$
$$c_n = 0$$

Now , we can use this to compute the matrix X_n, denoting genotype fraction in the n^{th} generation.

$$X_n = M \cdot X_{n-1}$$

,where

$$X_n = \begin{pmatrix} a_n \\ b_n \\ c_n \end{pmatrix} \quad ; \quad X_{n-1} = \begin{pmatrix} a_{n-1} \\ b_{n-1} \\ c_{n-1} \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Thus, we get;

$$X_n = M^n \cdot X_0$$

To compute this for large n's , we diagonalize M; let λ be a eigenvalue of M and v be its corresponding eigenvector. Then,

$$\det(M-\lambda\cdot I_3)=\begin{bmatrix}1-\lambda & \frac{1}{2} & 0\\ 0 & \frac{1}{2}-\lambda & 1\\ 0 & 0 & 0-\lambda\end{bmatrix}=(1-\lambda)\cdot(\frac{1}{2}-\lambda)\cdot(-\lambda)=0$$

So, the eigenvalues are 1, $\frac{1}{2}$ and 0 and its corresponding eigenvectors are ,

$$\lambda_1=1;\qquad M-\lambda_1\cdot I_3=\begin{pmatrix}0 & \frac{1}{2} & 0\\ 0 & \frac{-1}{2} & 1\\ 0 & 0 & -1\end{pmatrix},v_1=\begin{pmatrix}1\\0\\0\end{pmatrix}$$

$$\lambda_2=\frac{1}{2};\qquad M-\lambda_2\cdot I_3=\begin{pmatrix}\frac{-1}{2} & \frac{1}{2} & 0\\ 0 & -1 & 1\\ 0 & 0 & \frac{-1}{2}\end{pmatrix},v_2=\begin{pmatrix}1\\-1\\0\end{pmatrix}$$

$$\lambda_3=0;\qquad M-\lambda_3\cdot I_3=\begin{pmatrix}1 & 0 & 0\\ 0 & \frac{1}{2} & 1\\ 0 & 0 & 0\end{pmatrix},v_3=\begin{pmatrix}1\\-2\\1\end{pmatrix}$$

Thus , we get $M=P\cdot D\cdot P^{-1}$, where D is the diagonalized matrix and P is obtained by combining the 3 eigenvectors of M

$$D=\begin{pmatrix}1 & 0 & 0\\ 0 & \frac{1}{2} & 0\\ 0 & 0 & 0\end{pmatrix};\quad P=\begin{pmatrix}1 & 1 & 1\\ 0 & -1 & -2\\ 0 & 0 & 1\end{pmatrix};\quad P^{-1}=\begin{pmatrix}1 & 1 & 1\\ 0 & -1 & -2\\ 0 & 0 & 1\end{pmatrix}$$

Therefore ,

$$X_n=M^n\cdot X_0=P\cdot D^n\cdot P^{-1}\cdot X_0=\begin{pmatrix}1 & 1 & 1\\ 0 & -1 & -2\\ 0 & 0 & 1\end{pmatrix}\cdot\begin{pmatrix}1^n & 0 & 0\\ 0 & (\frac{1}{2})^n & 0\\ 0 & 0 & 0^n\end{pmatrix}\cdot\begin{pmatrix}1 & 1 & 1\\ 0 & -1 & -2\\ 0 & 0 & 1\end{pmatrix}\cdot\begin{pmatrix}a_0\\b_0\\c_0\end{pmatrix}$$

$$\implies X_n=\begin{pmatrix}1 & 1-(\frac{1}{2})^n & 1-(\frac{1}{2})^{n-1}\\ 0 & (\frac{1}{2})^n & (\frac{1}{2})^{n-2}\\ 0 & 0 & 0\end{pmatrix}\cdot\begin{pmatrix}a_0\\b_0\\c_0\end{pmatrix}=\begin{pmatrix}a_0+b_0\cdot(1-(\frac{1}{2})^n)+c_0\cdot(1-(\frac{1}{2})^{n-1})\\ \frac{b_0}{2^n}+\frac{c_0}{2^{n-2}}\\ 0\end{pmatrix}$$

Thus , we can predict the genotypic population density in the n^{th} generation . We also note that when $n \longrightarrow \infty$,

$$X_n \longrightarrow \begin{pmatrix} a_0 + b_0 + c_0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

2 GENE INTERACTION

Gene interaction is the process through which various genes or genetic elements can interact with one another to affect an organism's features, phenotypes, or risk of contracting diseases. Genes frequently cooperate with one another rather than acting separately, resulting in intricate relationships that let an organism function as a whole.

2.1 Gene-Gene Interactions

Gene-gene interactions, also known as genetic interactions, occur when the effect of one gene on a trait or phenotype is modified by another gene or set of genes.These interactions can be additive, multiplicative, or epistatic.

- Additive Interactions: Additive interactions are a type of gene-gene interaction where the combined effect of two or more genes on a trait or phenotype is equal to the sum of their individual effects. Example: Gene interactions to determine Human Eye Colour.
- Multiplicative Interactions: Multiplicative gene interactions, also known as synergistic, occur when the combined effect of two or more genes on a trait or phenotype is greater than the sum of their individual effects.
- Epistatic Interactions: It is a type of gene-gene interaction in which the effect of one gene (modifier gene) masks or modifies the effect of another gene (target gene) in determining a particular phenotype or trait. Example: Skin Colour in Rabbits is influenced by the interaction between the Extension (E) gene and the Agouti (A) gene. The E gene determines whether pigment production will occur, while the A gene influences the distribution of the pigment in the hair shaft. The interaction between these two genes can lead to various coat colors, including solid, agouti, and other patterns.

3 DISEASE SUSCEPTIBILITY

The chance or vulnerability of an individual to contract a specific disease is referred to as disease susceptibility. It is affected by a number of lifestyle, environmental, and hereditary variables. For the purpose of identifying risk factors, creating preventive measures, and offering individualised healthcare, understanding disease susceptibility is essential. Genetic variations can contribute to disease susceptibility by influencing an individual's response to environmental triggers or by directly affecting biological pathways involved in disease development.Thus , studying family history of a genetic disorder is really crucial. We will briefly discuss one hereditary

disorder(Haemophilia A) , and one mutational disorder(Sickle Cell Anaemia) .

3.1 Haemophilia

Hemophilia is an inherited bleeding disorder characterized by a deficiency or abnormality in certain clotting factors in the blood, specifically clotting factors VIII (hemophilia A) or IX (hemophilia B). Hemophilia is typically inherited in an X-linked recessive manner, that is it is caused due to change in gene present in X chromosome, and the disease symptoms occur only when all X chromosomes of an individual have been changed.

Pedigree analysis is a technique used in genetics to examine how traits or diseases are passed down across families over many generations. It entails creating a family tree or pedigree chart that shows the connections between family members and chronicles the passing down of traits or diseases.

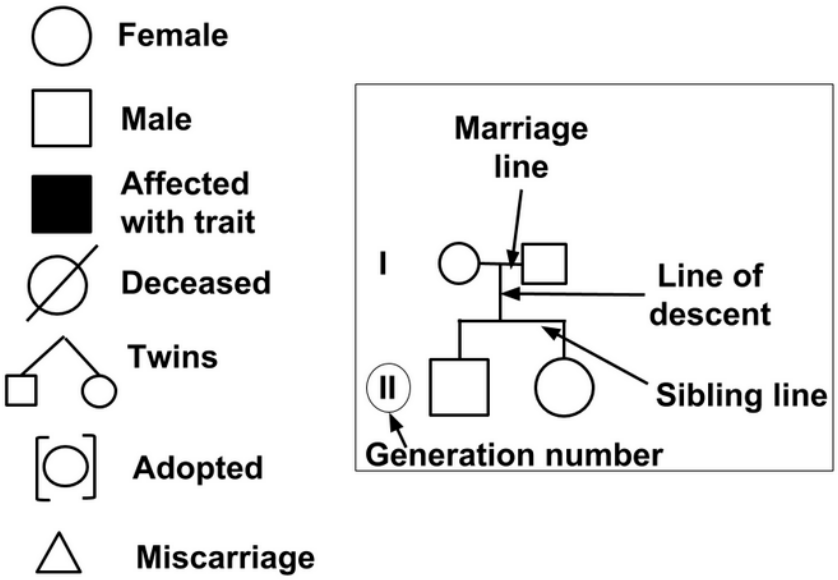

Figure 1: Common Symbols in a Pedigree Tree

We will go through a case study of Family Tree of Haemophilia A .

Consider the below family tree, we see that in G_0(Gen 0), the woman has a carrier of haemophilia. She isn't having the disorder, as only one allele is changed. Consider the following cases:

Male/Female	XX		X_cX		X_cX_c	
XY	XX	XX	X_cX	XX	X_cX	X_cX
	XY	XY	X_cY	XY	X_cY	X_cY
XY	XX_c	XX_c	X_cX_c	XX_c	X_cX_c	X_cX_c
	XY	XY	X_cY	XY	X_cY	X_cY

Table 2: Genotypes of Offsprings in Haemophilia

Male/Female	XX	X_cX	X_cX_c
XY	HF = 1/2 HM = 1/2 CF = 0 IF = 0 IM = 0	HF = 1/2 HM = 1/4 CF = 1/4 IF = 0 IM = 1/4	HF = 0 HM = 0 CF = 1/2 IF = 0 IM = 1/2
X_cY	HF = 0 HM = 1/2 CF = 1/2 IF = 0 IM = 0	HF = 1/4 HM = 1/4 CF = 1/4 IF = 1/4 IM = 1/4	HF = 0 HM = 0 CF = 0 IF = 1/2 IM = 1/2

Table 3: Probabilities of Disease Susceptibility in Offsprings during Haemophilia
Abrevations used: H - Healthy, C - Carrier, I - Infected, M - Male, F - Female

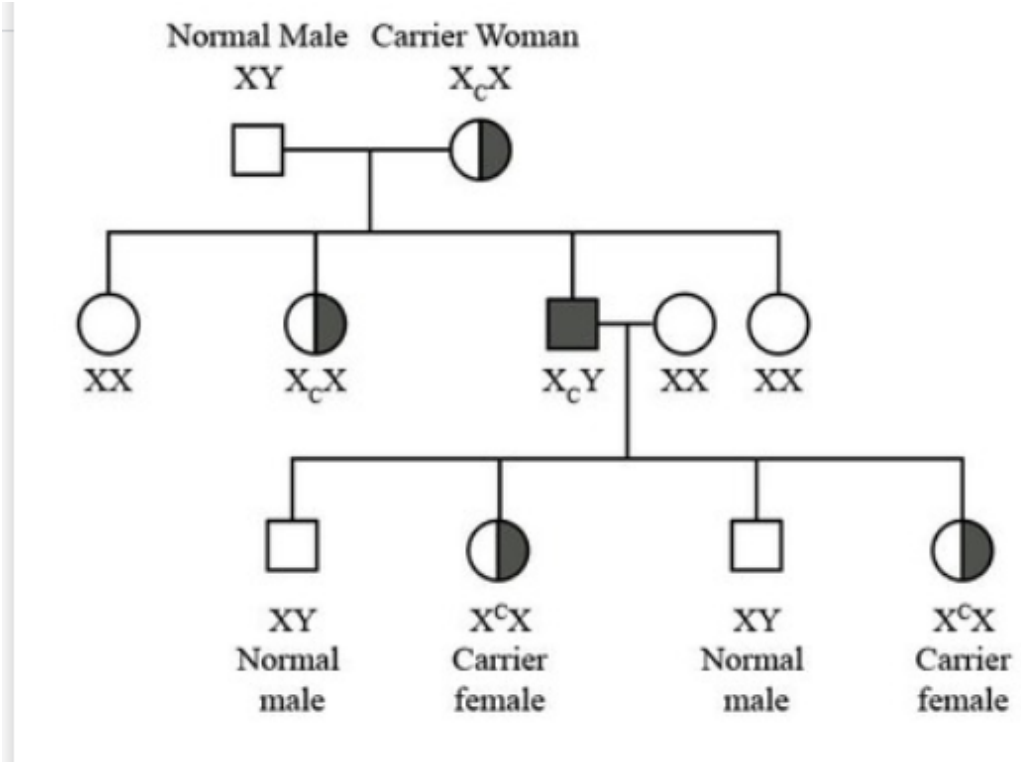

Figure 2: Pedigree Chart of Haemophilia

3.2 Sickle Cell Anaemia

Sickle cell anemia is a genetic blood disorder characterized by the presence of abnormal hemoglobin, called hemoglobin S (HbS), in red blood cells. Due to a nucleotide error that results in the creation of aberrant beta-chains in haemoglobin S, the single amino acid substitution of valine for glutamic acid in the beta-chain that causes sickle cell anaemia. In the deoxygenated state, abnormal haemoglobin chains form polymers, resulting in the distinctive sickle cells.

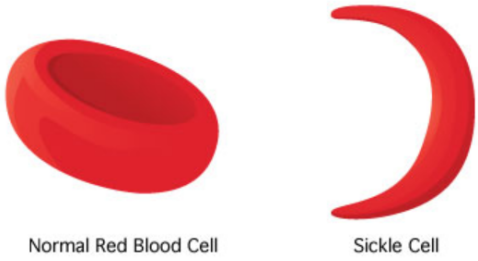

Figure 3: Comparison of a Normal RBC and a Sickle Cell

Sickle cell anemia is inherited in an autosomal recessive manner, that is, both copies of the HBB gene must carry the mutation for the disease to manifest.

Hemoglobin is responsible for carrying oxygen in red blood cells. In sickle cell anemia, the abnormal hemoglobin S causes red blood cells to become rigid and crescent-shaped, resembling sickles. These sickle-shaped cells can clump together and obstruct blood vessels, leading to reduced oxygen supply and tissue damage. This is really common in Africa. Because of the sickle shaped cells , people who have sickle cell anaemia don't get Malaria.

We can see the gene-gene network of HBB gene below.

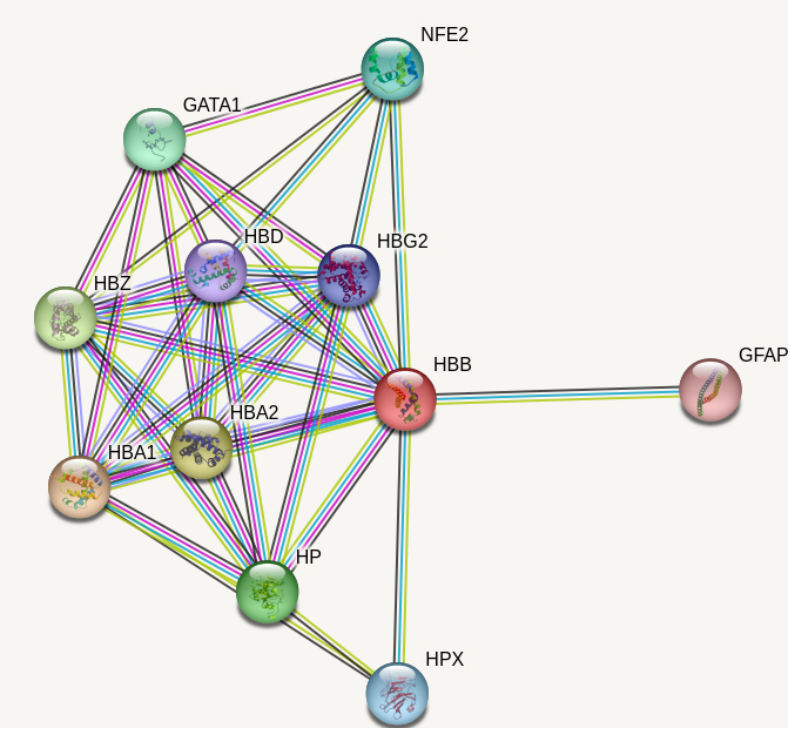

Figure 4: Gene-Gene Network of HBB gene (source)

4 MARKOV CHAINS

A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. Informally, this may be thought of as

"What happens next depends only on the state of affairs now."

Markov Chains in general are dependent on two key pieces of information:

- **Transition Matrix:** Denoted as "P," This NxN matrix represents the probability distribution of the state's transitions. The sum of probabilities in each row of the matrix will be one, implying that this is a stochastic matrix.

  Note that a directed, connected graph can be converted into a transition matrix. Each element in the matrix would represent a probability weight associated to an edge connecting two nodes.

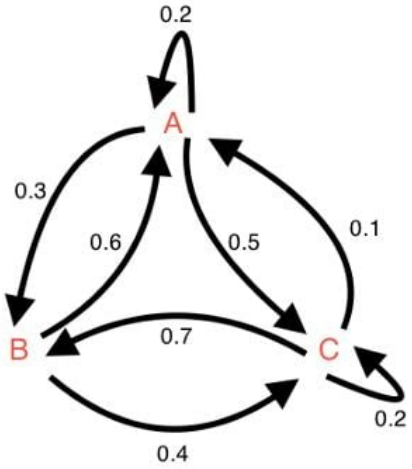

Figure 5: This graph outlines the probability associated with moving from one state to another. For example, there is a 60 percent chance to move from state B to state A.

This is the Transition matrix representation of the below graph structure.

```
      +-----+-----+-----+
      |  A  |  B  |  C  |   - Represents the network above
+-----+-----+-----+-----+   - NxN transition matrix
|  A  |  .2 |  .3 |  .5 |   - element hold probabilities
+-----+-----+-----+-----+   - row sum of probabilities = 1
|  B  |  .6 |  0  |  .4 |   - .3 is the probability for state A
+-----+-----+-----+-----+     to go to state B
|  C  |  .1 | .7  |  .2 |   - .7 is the probability for state C
+-----+-----+-----+-----+     to go to state B
```

- **Initial state Vector:** Denoted as "S," this Nx1 vector represents the probability distribution of starting at each of the N possible states. Every element in the vector represents the probability of beginning at that state.

Markov chains are essential tools in understanding, explaining, and predicting phenomena in computer science, physics, biology, economics, and finance. Now will study an application of linear algebra. You will see how the concepts we use, such as vectors and matrices, get applied to a particular problem.

4.1 Population analysis

4.1.1 Leslie Matrix

In applied mathematics, the Leslie matrix is a discrete, age-structured model of population growth that is very popular in population ecology. The Leslie matrix (also called the Leslie model) is one of the most well-known ways to describe the growth of populations (and their projected age distribution), in which a population is closed to migration, growing in an unlimited environment, and where only one sex, usually the female, is considered.

The below matrix L is called a Leslie Matrix, in general if we have a population with n classes of equal duration, L will be an $n \times n$ matrix with the following structure:

$$L = \begin{pmatrix} b_1 & b_2 & 0 & \dots & b_n \\ s_1 & 0 & 0 & \dots & 0 \\ 0 & s_2 & 0 & \dots & 0 \\ 0 & 0 & s_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_{n-1} \quad 0 \end{pmatrix}$$

Here $b_1, b_2, \dots$ are the birth parameters (b_i = the average numbers of females produced by each female in class i) and $s_1, s_2, \dots$ are the survival probabilities (s_i = the probability that a female in class i survives into class $i+1$).

Multiplying a matrix with its state vector gives the matrix in the next state just like a Markov chain. In this way, we can predict one or more steady states where this model can reach.

5 PREY-PREDATOR MODEL

Prey-Predator (also known in literature as Lotka-Volterra model) is a popular model to study dynamics of a system consisting of two antogonists, in this case rabbits (prey) and foxes (predator).

The dynamics of the sytem are determined by interactions within and between the prey and predator populations. The intra-species interactions are (natural) birth and (natural) death rates, while inter-species interactions are the predation of prey (i.e. predator 'eats' prey for its survival!). Let X denote the population size of prey and Y denote the popluation size of predator.

For the population dynamics of the prey: prey replicates at a rate that is controlled by abundance of the natural resources (rabbits need grass); we assume that these natural resources are abundant and remain at the same level throughout. Prey might die of natural causes (old age) or is eaten by predator. Thus the dynamics are reasonably modeled as:

$$\frac{dX}{dt} = \alpha X - \beta XY$$

For the population dynamics of predator: population of predator is expected increase linearly with its own size, and also on the population size of prey (since it needs prey as food). The natural death rate of the population depends on its own population size. Thus prey population size dynamics may be modeled as:

$$\frac{dY}{dt} = \gamma XY - \delta Y$$

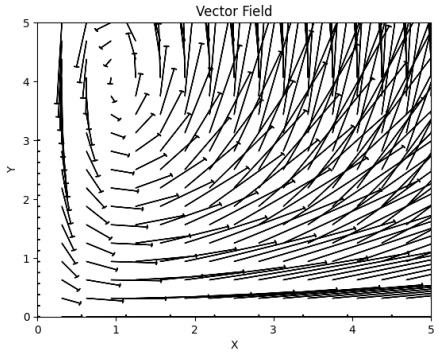

Figure 6: In the vector field plot, the arrows indicate the direction and magnitude of the rate of change of the prey and predator populations at different points in the XY plane. By examining the direction of the arrows, you can qualitatively understand the behavior of the system.

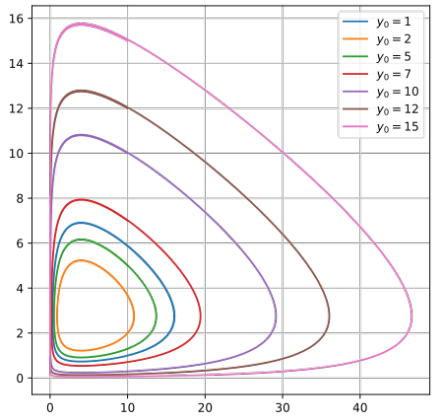

Figure 7: Phase-space plot for the predator-prey problem for various initial conditions of the predator population.

Clearly the dynamics of the model are dependent on the four positive constants α, β, γ, and δ, which are to be inferred from the feild data.

But this model also has its limitations:

- Linearity assumption: Linear algebra is based on the assumption of linearity, which may not hold true in all ecological systems.
- Lack of spatial considerations: Linear algebra treats populations as homogeneous entities without considering spatial heterogeneity.
- Simplified assumptions: Linear algebra models often rely on simplifying assumptions to facilitate mathematical tractability.
- Lack of feedback mechanisms: Prey-predator interactions involve feedback mechanisms, where changes in prey and predator populations influence each other.

The modern side of the research in the field of Lotka-Volterra models is under "Extinction Dynamics" and "Evolving Networks" [1], which is way beyond the scope of our project.

6 CPG ISLANDS

Before we jump into CpG islands we need to know a few things:

6.1 DNA Methylation

DNA Methylation, also called Dmeth in short, is an epigenetic modification of DNA that happens in all cells of your body. DMeth levels across the genome

have been successfully used to predict all kinds of diseases from [4] cancers to [10] Alzheimer's, your [2] biological age, and even your [6] time to death! The problem is that we still don't know what causes the changes in DMeth and which part affects what. Thus we engineer smarter features using DNA methylation data and subsequently use them in predicting cancer status (more about that later). In particular, we are looking for genomic features that underlie DNA Methylation states at different DNA sites.

6.2 HMM (Hidden Markov Chains)

Hidden Markov Model (HMM) is a statistical model that is used to describe the probabilistic relationship between a sequence of observations and a sequence of hidden states. It is often used in situations where the underlying system or process that generates the observations is unknown or hidden, hence it got the name "Hidden Markov Model". It is used to predict future observations or classify sequences based on the underlying hidden process that generates the data.

An HMM consists of two types of variables: hidden states and observations.

- The hidden states are the underlying variables that generate the observed data, but they are not directly observable.
- The observations are the variables that are measured and observed.

(Above is just a brief introduction of HMMs)

More in-depth information of Hidden Markov Chains is available in these links, links, and links.

6.3 THE CpG Island Model demonstration

So coming back to our main topic...

CpG islands (or CG islands) are regions with a high frequency of CpG sites. Though objective definitions for CpG islands are limited, the usual formal definition is a region with at least 200 bp (bond pairs), a GC percentage greater than 50

So we will move forward with this topic taking an example and replicating DMeth.

We have simulated a similar project (refer to the Jupyter notebook attached) which performs Logistic regression to analyze the DMeth.

The goal of this section of the project is to predict DNA methylation state (Beta 0 = unmethylated, Beta 1 = methylated) at any given CpG site. DNA samples are collected from more than 400 individuals and averaged to calculate the methylation state at each CpG site.

6.3.1 Methodology

Here's a breakdown of the steps involved:

1. **Data Loading**: The code starts by loading the training data from the "train.csv" file into a pandas DataFrame using the pd.read_csv() function.
2. **Data Preprocessing**: The code then performs some preprocessing steps on the data. First, it selects the relevant features for analysis and the "Beta" values. It uses the pd.get_dummies() function to convert categorical values into numerical values. The "Beta" values are appended to the feature matrix.
3. **Feature Analysis**: The code calculates the mean of the "Beta" values for each unique value in the "Island" feature. This provides an overview of the relationship between CpG islands and the "Beta" values.

4. **Data Split and Model Training**: The feature matrix X is separated from the target variable y, which contains the "Beta" values. The Logistic regression model is then trained on the training data using the LogisticRegression class from the scikit-learn library.

5. **Model Evaluation**: The code calculates the Area Under the ROC Curve (AUC) using the roc_auc_score() function from scikit-learn. AUC is a commonly used metric for evaluating binary classification models. Additionally, the code generates a classification report using the classification_report() function to provide detailed performance metrics such as precision, recall, and F1-score.

6. **Test Data Prediction**: The code loads the test data from the "test.csv" file and preprocesses it in the same way as the training data. The Logistic regression model is then used to predict the "Beta" values for the test data.

7. **Result Export**: The predicted "Beta" values are saved in the "solution.csv" file along with the other columns from the test data.

6.3.2 Results and Interpretation

By calculating the mean "Beta" value for each unique value in the "Island" feature, the code provides an overview of the average methylation levels associated with different types of CpG islands. This analysis implies that different CpG island types may exhibit varying levels of methylation. Further exploration and analysis of these associations can provide insights into the relationship between CpG islands and methylation levels.

The Logistic regression model trained on the provided training data is used to predict the methylation levels ("Beta" values) for the test data. This suggests that the model has learned patterns from the training data and can generalize its predictions to unseen data. However, the specific accuracy or quality of these predictions is not mentioned in the code.

Overall, this code provides a simple approach to analyze CpG islands using Logistic regression. It leverages the relation of CpG islands to specific locations and associated methylation levels to make predictions. The interpretation of the results and the insights gained depend on the specific data and context in which the analysis is applied.

GLOSSARY

Logistic regression A statistical model used to analyze the relationship between a dependent variable and one or more independent variables. 2, 3, 15, 16

ROC A ROC (Receiver Operating Characteristic) curve is a graphical representation used in binary classification tasks to assess the performance of a model or classifier. It plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds.. 2, 16

7 LOGISTIC MAPS

The logistic map is a quadratic map, i.e., $f(x)$ is a quadratic polynomial. It shows a particularly interesting phenomenon of 'deterministic chaos', i.e., a deterministic map showing apparently random behavior. The logistic map has a single parameter, named α, and is given by:

$$f(x) = \alpha x(1-x)$$

When $0 \leqslant \alpha \leqslant 4$, the map takes an input $0 \leqslant x \leqslant 1$ to give an output in the same range. For a particular value of the parameter α, we want to find the behavior of the map. For such systems, a plot of x_n vs x_{n+1} is called a phase plot and is an important tool for visualizing and analyzing such systems.

7.1 Population Dynamics and Spatial Heterogeneity

The logistic map is a mathematical model used to predict population growth and demonstrate chaotic behavior.

Guppies example: It applies to a scenario involving an aquarium with guppies, where the question arises: How many guppies can be expected next month if they are left to breed freely?

The logistic map considers the initial population and its proximity to the maximum capacity of the aquarium. If the population is far from the maximum capacity, indicating sufficient resources, it is likely to increase. Conversely, if the population is close to the maximum capacity, resources will be limited, leading to a decrease in population.

The logistic map introduces the variable x to represent the proportion of the tank's maximum capacity occupied by guppies. x varies between 0 (no guppies) and 1 (maximum capacity of 100 guppies).

The formula $rx(1-x)$ estimates the next month's population size, where r is the growth rate determined by factors such as birth and death rates.

One of the key observations is that when the value of r increases, above the value of 3, the population change settles into a periodic pattern, but beyond $r = 3.56995$ (approx), the population may never settle down into a predictable pattern. Moreover, two different starting values of x can lead to wildly different predictions of the future, even if they are very, very close.

7.2 Logistic regression and Disease Mapping on Large Moving Populations

Spatial analysis of disease risk, or disease mapping, typically relies on information about the residence and health status of individuals from the population under study. However, residence information has its limitations because people are exposed to numerous disease risks as they spend time outside of their residences [5]

Thanks to the widespread use of mobile phones and GPS-enabled devices, it is becoming possible to obtain a detailed record of the movement of human populations. The availability of movement information opens up an opportunity to improve the accuracy of disease mapping.

The 'alpha' in the logistic equation has its own form in disease mapping, formally known as the 'r' factor.

COVID-19 example:

When the total number of cases is plotted against a logarithmic scale, we can visualize the exponential growth of COVID-19 cases. Using simple regression analysis (a statistical method to obtain a mathematical function using the least-square-error method), a model is fitted through the WHO's data in the log scale. It turns out that the total number of COVID-19 cases on a daily basis can be easily modeled using the exponential growth equation (equation 1 given below):

$$n_{i+1} = r \cdot n_i$$

According to the model fitted to the WHO's data for cases outside China, the growth rate (r) is approximately 1.18564. However, since this model is exponential, it diverges to infinity over time, which is physically incorrect. As more people are infected, they are surrounded by people who are already infected or have developed

immunity. So there is a saturation point in the infected population. Typically, in order to predict the growth, another factor is multiplied on the right-hand side of this exponential growth equation:

$$n_{i+1} = r \cdot n_i \cdot (1 - \frac{n_i}{N})$$

Replacing n_i/N with a variable x, we get the logistic map equation. Now, the logistic map is a very simplified yet powerful way to model the spread of infections like COVID-19.

8 PCA

Principal Component Analysis is a dimensionality reduction method that is often used to reduce dimensionality of high data sets by transforming,

ADVANTAGES OF PCA: PCA is a very important tool for analysing high dimensional datasets, such as gene expression measurements where some of the biggest challenges stem from the vast amount of data.

Gene expression samples can have 10000+ measurements per sample and performing statistical analysis on these samples using regular methods such as linear regression is not feasible due to the high volume of data.

Thus, PCA which can reduce the data to a 2-D or 3-D plot makes the visualization, exploration and analysis of the data actually possible.

REPLICATING PCA: PCA is a comparitively simple process which can be replicated anywhere from a simple python script, to massive machine learning datacenters, In general PCA can be performed as follows:

1. Compute the mean-centered data matrix Z by subtracting the mean of each column from the corresponding values in X.
2. Compute the covariance matrix C of Z.
3. Perform eigendecomposition on C to obtain its eigenvectors (principal components) V and corresponding eigenvalues λ.
4. Sort the eigenvalues in descending order and select the top k eigenvectors corresponding to the largest eigenvalues, where k is the desired number of dimensions for the reduced representation.
5. Transform the data matrix Z into a lower-dimensional representation Y by multiplying Z with the selected eigenvectors.

8.1 Example: PCA for capturing global immune response patterns

The mathematical equation for PCA involves computing the eigenvectors and eigenvalues of the covariance matrix of the data. Given a data matrix X, where each row represents a sample and each column represents a variable (in this case, the measurements of IgM and IgG), [3]

8.2 Example: PCA for analyzing genetic variation

In the context of analyzing genetic variation, PCA can be used to summarize and visualize patterns of genetic differentiation among different populations. The mathematical equation for PCA remains the same as described above.

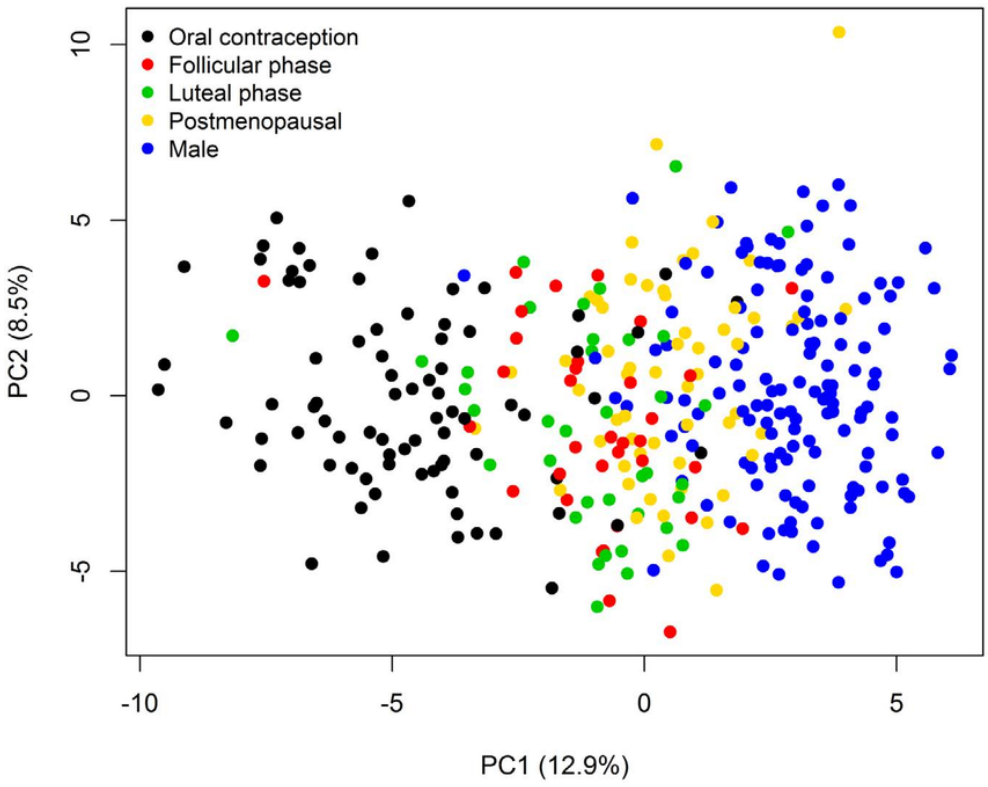

Figure 8: Principal component analysis (PCA) plot of NESDA discovery cohort samples. The first two principal components (PCs) are plotted and colored according to sex and female hormonal status. PCA was performed using all analyte data. Percentage of variation accounted for by each principal component is shown in brackets with the axis label.

In this case, the data matrix X would typically consist of genetic markers or genotypes, such as single nucleotide polymorphisms (SNPs), across individuals from different populations. The PCA analysis allows for the identification of principal components that capture the major sources of genetic variation across populations. [7]

9 CONCLUSION

USE OF LINEAR ALGEBRA: Use case of Linear Algebra in the Project: Our project unconventionally approaches the Application of Linear Algebra in Genetics, from mathematical point of view. This approach lets the readers understand how a given Mathematical model can be applied in various fields including Biology. Some of the major mathematical concepts are listed above like application of Markov Chains, Principal Component Analysis (PCA), Statistical Regression.

In summary, this project has shed light on the intricate relationships within biological systems and provided valuable insights into disease susceptibility, population dynamics, genetic variation, and immune response patterns. By leveraging various mathematical models and analytical techniques, we have deepened our understanding of gene interactions and their implications. These findings contribute to the broader field of biomedical research, offering potential avenues for disease prevention, diagnosis, and treatment.

REFERENCES

[1] M. Ausloos, *The logistic map and the route to chaos: From the beginnings to modern applications*, Springer Science & Business Media, 2006.

[2] C. G. Bell, R. Lowe, P. D. Adams, A. A. Baccarelli, S. Beck, J. T. Bell, B. C. Christensen, V. N. Gladyshev, B. T. Heijmans, S. Horvath, T. Ideker, J.-P. J. Issa, K. T. Kelsey, R. E. Marioni, W. Reik, C. L. Relton, L. C. Schalkwyk, A. E. Teschendorff, W. Wagner, K. Zhang, and V. K. Rakyan, *Dna methylation aging clocks: challenges and recommendations*, Genome Biology, 20 (2019), p. 249.

[3] C. Fesel and A. Coutinho, *From igm to igg: Direct amplification and switch of antibody response upon lymphocytic choriomeningitis virus infection*, European Journal of Immunology, 39 (2009), pp. 1489–1502.

[4] M. Kulis and M. Esteller, *Dna methylation and cancer*, Advances in Genetics, 70 (2010), pp. 27–56.

[5] A. L. Lloyd, *The coupled logistic map: a simple model for the effects of spatial heterogeneity on population dynamics*, Journal of Theoretical Biology, 173 (1995), pp. 217–230.

[6] A. T. Lu, A. Quach, J. G. Wilson, A. P. Reiner, A. Aviv, K. Raj, L. Hou, A. A. Baccarelli, Y. Li, J. D. Stewart, E. A. Whitsel, T. L. Assimes, L. Ferrucci, and S. Horvath, *Dna methylation grimage strongly predicts lifespan and healthspan*, Aging, 11 (1945), pp. 303–327.

[7] G. McVean, *A genealogical interpretation of principal components analysis*, PLoS Genetics, 5 (2009), p. e1000686.

[8] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, *Dimensionality reduction using genetic algorithms*, IEEE transactions on evolutionary computation, 4 (2000), pp. 164–171.

[9] J. B. Reece, N. Meyers, L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, R. B. Jackson, B. J. Cooke, and N. A. Campbell, *Campbell Biology*, Pearson, 2015.

[10] A. S. Yokoyama, J. C. Rutledge, and V. Medici, *Dna methylation alterations in alzheimer's disease*, Environ Epigenet, 4 (2017).

[11] G. Zhang, Z. Dai, and X. Dai, *C-rnncrispr: Prediction of crispr/cas9 sgrna activity using convolutional and recurrent neural networks*, Computational and structural biotechnology journal, 18 (2020), pp. 344–354.

[12] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, *Convolutional recurrent neural networks: Learning spatial dependencies for image representation*, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 18–26.