An Investigation of Racial Profiling in an Airport Setting

Samantha Shepler
sshepler08@winona.edu
Advisor: Chris Malone
Winona State University

Abstract

Racial profiling can be defined as any actuarial method that conditions an individual's prior probability of criminal behavior explicitly on his or her race, ethnicity, nationality or religion (Press, 2010). Many questions have been raised whether the method of racial profiling is ethical, but is it even effective? The use of racial profiling within airport security is investigated here through simulation. The goal of this project is to determine whether or not strong racial profiling is more effective than simple uniformly random sampling in catching a potential terrorist.

Background

The idea for this project came from an article written by Willam Press called "To Catch a Terrorist: Can Ethnic Profiling Work?" This article describes two basic methods for airport screening, uniform random sampling and importance sampling.

Uniform Random Sampling: Randomly selecting passengers to go through secondary screening without using their prior probability of being a security threat.

Importance Sampling: Using prior probabilities when selecting passengers to go through secondary screening. For Example, if we think a specific group of people is twice as likely to be a security threat we will pull those passengers out twice as often for a secondary screening .

These passengers are thought to be twice as likely to be a security threat. This could be based on many different things including age, race, or name.

Press (2010), derives the average number of checkpoints a security threat can get through without being caught. Mathematically, the averages between the different sampling methods are equivalent.

Uniform Sampling

$$\mu = \sum_{i=1}^{N} \frac{p_i}{M/N} = \frac{N}{M}$$

Importance

$$\mu = \sum_{i=1}^{N} \frac{p_i}{M p_i} = \frac{N}{M}$$

N equals numbers of passengers to pass through security checkpoint, p_i is the individual prior probability for each passenger, and M is the amount of passengers based on limited resources that can have a secondary screening.

Simulation

To do the following simulation we created a function in R for each sampling method.

Step 1: chind[1:num.passengers, c(0,0,0,1,...,0), prob.st=c(.05,.05,.01,.05,...,.05)]

- Create matrix of data. Since there is no accessible data on this topic we had to create the data. The matrix included a ID number, randomly assigned security threat status, and probability of being a security threat.

Step 2: for(i in 1:100) {r=sample(1:dim(newx)[1], p, replace=FALSE, prob.st)

- Create screening process. For one iteration, the data is run through using the specified method of either uniform or importance sampling. If a security threat is selected for secondary screening it is removed from the data. The data runs through this process until all security threats have been removed.

Step 3: for(j in 1:num.iter) {code from step 2}

- Repeat. To get enough data we wanted to be able to run many iterations.

Step 4: return(avg.sim)

- Report results. The function returns the average checkpoint the security threats were caught for each iteration and the average checkpoint over all of the iterations.

Comparing Means

This example assumes 10,000 passengers a day over a year where 1 out of every 1000 of the passengers is a security threat. The level of resources dedicated is varied in this example.

When resources allow to have 1 out of every 7 passengers screened

Method	Mean Checkpoint
Theory	7
Importance Sampling	7.035
Uniform Sampling	6.848

When resources allow to have 1 out of every 10 passengers screened

Method	Mean Checkpoint
Theory	10
Importance Sampling	10.01
Uniform Sampling	10.03

It is easy to see that the means are similar across the different screening methods for both levels of resources. These means also agree with the mathematically derived results and the graphs reiterate these findings.

When resources allow to have 1 out of every 7 passengers screened

When resources allow to have 1 out of every 10 passengers screened

Comparing Distributions

When looking at the previous plots, it seems that even though the means for the different screening methods look similar the distributions of the data may differ. The plots show that the distributions for the different screening methods follow very similar trends where they peak before the mean and are slightly skewed right. It is very interesting that when resources allow 1 out of 7 passengers to be screened the importance sampling method seems to be more effective at catching security threats only until slightly before security checkpoint 7, then the uniform sampling method is better until all security threats are caught. This trend is also apparent when 1 out of 10 passengers can be screened but the screening methods switch effectiveness slightly before security checkpoint 10.

Additional Investigations

The previous examples showed the simulation provided results that seem to agree with the mathematically derived results but it is not realistic for an airport to be able to give a secondary screening to every 7th or 10th passenger. This example assumes 10,000 passengers a day over a year where 1 out of every 1000 of the passengers is a security threat and resources allow 1 out of 100 passengers to be screened.

Method	Mean Checkpoint
Theory	100
Importance Sampling	99.60
Uniform Sampling	103.26

This example shows that even with limited resources, the simulation still agrees with the mathematically derived results where when 1 out of 100 people can be given a secondary screening a security threat will on average be caught at the 100th checkpoint. While the data is much more spread out, the distribution of this example has similar trends to the previous example.

Conclusions

- The simulation has provided results which seem to agree with the mathematically derived results from Press. Importance sampling based on things like race, religion or name is no more effective at catching security threats in airports than uniform random sampling.

References

- Press, W. (2010). To Catch a Terrorist: Can Ethnic Profiling Work? Significance, 7(4), 164-167.
- Press, W. (2008). Strong Profiling is Not Mathematically Optimal for Discovering Rare Malfeasors. Proceeding of the National Academy of Sciences of the United States of America, 106(6), 1716-1719. doi:10.1073/pnas.0813202106.

Classical Cipher Protection

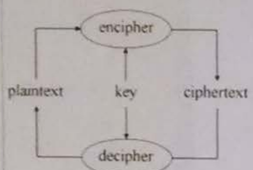
Anne Longlet
Faculty Advisor: Dr. Joyati Debnath, Mathematics and Statistics
Winona State University, Winona MN

Overview

- Pass information from one person to another without any outside party knowing what the information is
- A way to encrypt information is called a cipher
- The more difficult the cipher is to decrypt, the more protected information is
- There are ways to combine different types of ciphers to achieve more protected information
- The method of combining ciphers together results in a "Lock and Key," model

Introduction

- Cryptography's purpose is to restrict information to a group of people and keep the information hidden from third parties.



- Ways to encrypt a message are called ciphers.
- These ciphers use modular arithmetic and matrix operations.
- A message called plaintext is encrypted using an encipher.
- The key used to encrypt is the same used to decrypt.
- The common ciphers that have been in use for centuries are easily broken on their own.
- These ciphers are easy to break but if two are combined, the more difficult it is to break them.

Definitions of Ciphers

1. Shift ciphers: The letters of the alphabet are shifted a number, *n*. This is an algebraic system, not random.
2. Substitution ciphers: Similar to the shift cipher, except the assignments of the numbers requires no pattern. Even though it seems random, it is still a chosen configuration.
3. Playfair cipher: This cipher uses a 5x5 matrix of letters to encode a message that as been broken up into two letter pairs.
4. Block cipher: This cipher uses an nxn matrix with a non-zero determinant to encrypt a message

Methods

Lock and Key Method 1: Playfair Cipher and Shift Cipher

1	2	3	4	5	6	7	8	9	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

15	11	1	24	6
9	17	19	2	12
10	25	18	16	14
8	7	4	3	22
5	13	21	23	20

Plaintext: *WINONA STATEX*

1. Apply the Playfair cipher by splitting plaintext into pairs, find numbers using the table. Use matrix to encrypt the numbers and convert to letters using the table.

$$\begin{bmatrix} W \\ I \end{bmatrix} = \begin{bmatrix} 22 \\ 9 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix} = \begin{bmatrix} h \\ m \end{bmatrix} \qquad \begin{bmatrix} N \\ O \end{bmatrix} = \begin{bmatrix} 13 \\ 14 \end{bmatrix} = \begin{bmatrix} 20 \\ 25 \end{bmatrix} = \begin{bmatrix} u \\ z \end{bmatrix}$$

$$\begin{bmatrix} N \\ A \end{bmatrix} = \begin{bmatrix} 13 \\ 1 \end{bmatrix} = \begin{bmatrix} 21 \\ 11 \end{bmatrix} = \begin{bmatrix} v \\ l \end{bmatrix} \qquad \begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} 18 \\ 19 \end{bmatrix} = \begin{bmatrix} 4 \\ 18 \end{bmatrix} = \begin{bmatrix} d \\ s \end{bmatrix}$$

$$\begin{bmatrix} A \\ T \end{bmatrix} = \begin{bmatrix} 1 \\ 19 \end{bmatrix} = \begin{bmatrix} 19 \\ 18 \end{bmatrix} = \begin{bmatrix} t \\ s \end{bmatrix} \qquad \begin{bmatrix} E \\ X \end{bmatrix} = \begin{bmatrix} 5 \\ 23 \end{bmatrix} = \begin{bmatrix} 13 \\ 20 \end{bmatrix} = \begin{bmatrix} n \\ u \end{bmatrix}$$

2. Apply encryption matrix below.
3. Use the additive inverses mod 26 to add encryption.

$$\text{encryption matrix} = \begin{bmatrix} 8 & 20 & 21 & 4 & 19 & 13 \\ 12 & 25 & 11 & 18 & 18 & 20 \end{bmatrix}$$

$$\text{encryption matrix of additive inverses in } \mathbb{Z}_{26} = \begin{bmatrix} 18 & 6 & 5 & 22 & 7 & 13 \\ 14 & 1 & 15 & 8 & 8 & 6 \end{bmatrix}$$

$$\begin{bmatrix} 18 & 6 & 5 & 22 & 7 & 13 \\ 14 & 1 & 15 & 8 & 8 & 6 \end{bmatrix} = \begin{bmatrix} s & f & e & w & g & n \\ o & a & p & h & h & f \end{bmatrix}$$

4. Matrix of the letters matching the additive inverses is the ciphertext.
5. To decrypt it, one has to work backwards.
6. The person for whom the message is intended would have prior knowledge of how to decrypt it.
7. In this case the lock and key would be the Playfair matrix and the additive inverses.

Lock and Key Method 2: Substitution and Block Cipher

The lock is a substitution and key will a 4x4 encryption matrix *B*.
Plaintext: *WINONA STATEX*.

1. Find the letter substitution for the plaintext, convert to numbers.

PT	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CT	q	w	e	r	t	y	u	i	o	p	a	s	d	f	g	h	j	k	l	z	x	c	v	b	n	m
PT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
CT	16	22	4	17	19	24	20	8	14	15	0	18	3	5	6	7	9	10	11	25	23	2	21	1	13	12

$$\begin{bmatrix} W & O & S & T \\ I & N & T & E \\ N & A & A & X \end{bmatrix} \rightarrow \begin{bmatrix} v & g & l & z \\ o & f & z & t \\ f & q & q & b \end{bmatrix} \rightarrow \begin{bmatrix} 21 & 6 & 11 & 25 \\ 14 & 5 & 25 & 19 \\ 5 & 16 & 16 & 1 \end{bmatrix} = A$$

2. Use an encryption matrix *B* to encryption matrix *A*.
3. Reduce mod 26.
4. Find letter counterparts of the resulting numbers.

Methods, continued.

$$B = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 3 & 1 & 0 \\ 1 & 2 & 0 & 0 \\ 2 & 0 & 1 & 2 \end{bmatrix} \text{ and } A \cdot B = \begin{bmatrix} 21 & 6 & 11 & 25 \\ 14 & 5 & 25 & 10 \\ 5 & 16 & 16 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 3 & 1 & 0 \\ 1 & 2 & 0 & 0 \\ 2 & 0 & 1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 82 & 40 & 52 & 71 \\ 77 & 65 & 38 & 52 \\ 23 & 80 & 22 & 7 \end{bmatrix} = \begin{bmatrix} 4 & 14 & 0 & 19 \\ 25 & 13 & 12 & 0 \\ 23 & 2 & 22 & 7 \end{bmatrix}_{26} = \begin{bmatrix} e & o & a & t \\ z & n & m & a \\ x & c & w & h \end{bmatrix}$$

5. This last matrix is the encrypted message to be sent between two people.
6. The lock is the substitution and the key is the encryption matrix.
7. To decrypt it, one would just have to work backwards, in the following way:

$$B^{-1} = \frac{1}{3}\begin{bmatrix} 4 & -2 & 3 & -2 \\ -2 & 1 & 0 & 1 \\ 6 & 0 & 0 & -3 \\ -7 & 2 & -3 & 5 \end{bmatrix} = \begin{bmatrix} 10 & 8 & 1 & 8 \\ 8 & 9 & 0 & 9 \\ 2 & 0 & 0 & 25 \\ 15 & 18 & 25 & 19 \end{bmatrix}_{26}$$

$$A \cdot B \cdot B^{-1} = \begin{bmatrix} 4 & 14 & 0 & 19 \\ 25 & 13 & 12 & 0 \\ 23 & 2 & 22 & 7 \end{bmatrix}_{26} \cdot \begin{bmatrix} 10 & 8 & 1 & 8 \\ 8 & 9 & 0 & 9 \\ 2 & 0 & 0 & 25 \\ 15 & 18 & 25 & 19 \end{bmatrix}_{26}$$

$$= \begin{bmatrix} 437 & 500 & 479 & 519 \\ 378 & 517 & 25 & 617 \\ 395 & 328 & 198 & 885 \end{bmatrix} = \begin{bmatrix} 21 & 6 & 11 & 25 \\ 14 & 5 & 25 & 19 \\ 5 & 16 & 16 & 1 \end{bmatrix} = \begin{bmatrix} v & g & l & z \\ o & f & z & t \\ f & q & q & b \end{bmatrix} = \begin{bmatrix} W & O & S & T \\ I & N & T & E \\ N & A & A & X \end{bmatrix}$$

Note: the encryption matrix *B* must be invertible i.e. the determinant can be zero. Also, since this takes place in the integers mod26, there can be no non-integers in an matrix.

Ideas For Further Research

1. Linear Operators – The encrypting matrix would be a linear operator and also using concepts of injectivity, surjectivity, and bijectivity.
2. Eigenvalues/Eigenvectors – Using diagonalization, the matrix would be split into three square matrices. The matrix starting out is invertible, i.e. it cannot have a determinant of zero.
3. Homomorphic Encryption – encryption using the ideas that are the same as homomorphisms. Operations performed on a ciphertext, the result is an encrypted message, which when decrypted will give the original plaintext.

Conclusions

- Due to the Internet, classical cryptography is not very useful since computers could decrypt a relatively simple cipher very quickly
- Since nearly everyone uses the internet, the need to protect personal information has grown.
- Cryptography has become more scientific rather than creative
- The objective of cryptography has not changed
- While working in classical cryptography, combining two (or possibly more) ciphers works for keeping information restricted
- This is a concept that could cross over to modern cryptography

Contact

Anne Longlet
Winona State University
Mathematics and Statistics Department
alonglet07@winona.edu

References

1. Cryptography: Theory and Practice – Douglas R. Stinson
2. Introduction to Modern Cryptography – Jonathan Katz
3. Introduction to Cryptography: With Coding Theory
4. [illegible]
5. [illegible]
6. [illegible]
7. [illegible]

Acknowledgments

- Joyati Debnath
- Winona State University

Semiparametric Regression for Manufacturing Data

Kristin Mara

Co Researchers: Samantha Meadows and Rosemarie Roessel

ABSTRACT

One of the approaches to model the smooth function in a nonparametric model is to approximate it by adopting adequate basis functions. We approximated the smooth function by truncated polynomial basis with degree 2, which contains the polynomial basis and the splines constructed by knots. The function is then estimated by well-known methods such as ordinary least squares, penalized spline and linear mixed model regression. We propose our version of Bayesian penalized spline regression, which provides comparable results. The prior distribution is chosen to be "objective" so it will minimize the influence to the posterior distribution and maintain the advantages Bayesian statistics provided. The non-informative Jeffreys prior is adopted for the polynomial basis and the variance component in the model. The prior for the splines constructed by knots is elicited from the penalty term in the penalized likelihood. To ensure the posterior distributions are proper, we have to use an informative prior on the smoothing parameter. To achieve the goal of an "objective" prior for smoothing parameter, we use the effective degree of freedom for fit to determine the hyperparameter in the prior distribution. We use the Akaike Information Criterion (AIC) to compare those methods proposed through a simulated and a manufacturing data set.

SEMIPARAMETRIC REGRESSION

For semiparametric regression, we have $y = X_1\beta_x + X_2\beta_{dum} + Z\mu + \varepsilon$

$$\text{where } X_1 = \begin{pmatrix} 1 & x_1 & \cdots & x_1^p \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p \end{pmatrix}, \beta_x = (\beta_0, \beta_1, \ldots, \beta_p)',$$

$$Z = \begin{pmatrix} (x_1 - K_1)_+^p & (x_1 - K_2)_+^p & \cdots & (x_1 - K_k)_+^p \\ (x_2 - K_1)_+^p & (x_2 - K_2)_+^p & \cdots & (x_2 - K_k)_+^p \\ \vdots & \vdots & \ddots & \vdots \\ (x_n - K_1)_+^p & (x_n - K_2)_+^p & \cdots & (x_n - K_k)_+^p \end{pmatrix}, \mu = (\beta_{p+1}, \beta_{p+2}, \ldots, \beta_{p+k})',$$

$$\text{and } X_2 = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1j} \\ M_{21} & M_{22} & \cdots & M_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nj} \end{pmatrix}, \beta_{dum} = (\beta_{p+k+1}, \ldots, \beta_{p+k+j})',$$

where n is the number of data points we have, p is the degree of the polynomial we are using to fit our data, k is the number of knots in the model, K_i is the value at which the i^{th} knot is placed, and j is the number of dummy variables in the model.

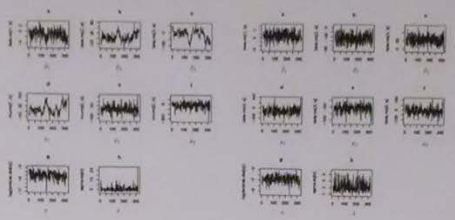
SIMULATED DATA

To begin our research we wanted to start with a data set that we knew how it was supposed to act, so we simulated a data set using

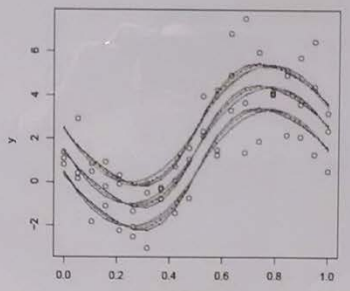
$$y_1 = f(time, op_1) + \varepsilon$$
$$y_2 = f(time, op_2) + \varepsilon$$
$$y_3 = f(time, op_3) + \varepsilon$$
$$\text{and } y = (y_1, y_2, y_3)$$

where $f(time, op_i) = i + 3 * sin(2 * time * \pi - \pi)$ and *time* is twenty even increments between 0 and 1. We used "time" as our continuous predictor variable, "op" (with j = 3 levels) as our categorical predictor, and "y" was our response variable. Thus our n = 60 and we decided to use a 2nd degree polynomial for this research. We used the adjusted R^2 to determine the number of knots that were appropriate for our model and found that we should use k = 3 knots for this data set.

We can see in the following plot that where one graph goes up, another one goes down to compensate for it. To improve this mix between β_x and μ, we turned to block sampling. To do this, we put β_x and μ in the same vector so that we could solve for them simultaneously. Doing this allows for fewer iterations due to rapid convergence, and it is easier to compute if we have a more complicated linear model.



Traceplot without block sampling (left), trace plot with block sampling (right)



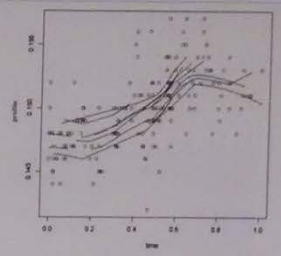
Plot of ordinary least squares (black solid), penalized spline (black dotted), linear mixed model (green), non-informative Bayesian penalized spline (blue), and informative Bayesian penalized spline (magenta)

Approach	AIC
Ordinary Least Squares	4.541667
Penalized Spline	4.536191
Linear Mixed Model	4.536197
Bayesian Penalized Spline (non-informative prior)	4.548832
Bayesian Penalized Spine (informative prior)	4.512195

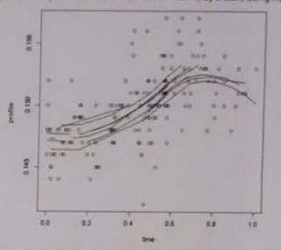
MANUFACTURING DATA

After working with the simulated data set, we finally had the tools to work with the manufacturing data. This new data concerns parts called "004 profiles", which are produced by stamping out metal material through a machine. Some parts of the machine become worn out after a certain period of time, resulting in the "004 profiles" being out of specification.

In this data set there are three variables. There is "profile", which is the measurement taken; "time", which is the time at which the measurement was taken; and there is also "operator" which is the ID of the operator that took the measurement. We considered "profile" to be our response variable, "time" was our continuous predictor variable, and "operator" is a categorical predictor variable. Thus we analyzed the data with the goal of developing the best prediction model. Using the adjusted R^2, we decided to use k = 5 knots for our models for the manufacturing data "004 profiles".



Plot of Bayesian penalized spline fit where b = 0.000047 with 18 operators using non-informative prior



Plot of Bayesian penalized spline fit where b = 0.000047 with 18 operators using an informative prior

Approach	AIC
Ordinary Least Squares	-6.933534
Penalized Spline	-6.936578
Linear Mixed Model	-6.936503
Bayesian Penalized Spline (non-informative prior)	-6.929622
Bayesian Penalized Spine (informative prior)	-6.929623

CONCLUSIONS

We found that when dealing with a semiparametric model for our simulated data, all of the methods gave approximately the same result. Looking between the non-informative and the informative prior for β_{dum}, we also saw very similar results. We believe that this is because of our small A value, which caused the variance $\frac{\sigma^2}{A}$ for our informative prior for β_{dum} to be very large. This in turn produced an almost uniform distribution, which was how our non-informative prior was classified, thus they gave very similar results. Therefore, the data provided sufficient information to overcome the influence of the priors selected.

FUTURE WORK

Some extensions of this project would be to look at additive models (combines two continuous functions), work with other bases (such as the radial basis function), and to look at hypothesis testing.

REFERENCES

[1] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin, *Bayesian data analysis*, CRC Press LLC., 2004.

[2] David Ruppert, Matthew P. Wand, and Raymond J. Carroll, *Semiparametric regression*, Cambridge UP., 2003.

Determining Future Genotype Proportions

Jaron Arbet and Shaun Miller

Faculty Advisor: Dr. Joyati Debnath

Mathematics and Statistics, Winona State University

Overview

- Goal: predict the genotype proportions of future generations
- Useful for limiting undesired genes, or increasing beneficial genes in future generations
- For example, these methods could be used to limit the occurrence of certain genes associated with diseases in future generations of a given population; or they could be used to increase genes in livestock and crops that would help maximize food production

Main Objective

- Using Markov Chain matrices to predict future genotype proportions of a population for both autosomal and X-linked genetic inheritance

Autosomal Inheritance

- Each parent can either pass down a dominant allele, a form of gene, (e.g. "G") or a recessive allele (e.g. "g"). If a dominant allele is paired with a recessive allele, the dominant phenotype will be expressed
- We chose to always have at least 1 parent with the dominant gene, thus we are simulating the extinction of an undesired recessive gene.
- Punnett Square Table below summarizes all possible genotype combinations of an offspring for some gene "G"

Parents' Genotype	Offspring Genotype	Probability
GG-GG	GG	1
	Gg	0
	Gg	0
GG-Gg	GG	.5
	Gg	.5
	Gg	0
GG-gg	GG	0
	Gg	1
	Gg	0

- The probabilities from the Punnett Square Table are put into a transformation matrix "A"

$$A = \begin{bmatrix} 1 & .5 & 0 \\ 0 & .5 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

- We use the following equation to predict future genotype proportions

$$X_n = A^n X_0 \text{ (Daddel, Genetics)}$$

- Where X_n and X_0 are the genotype proportions of the nth generation and beginning generation respectively.

Shortcut Method

- We use the following for computing A^n, when matrix A is diagonalizable:

$$A^n = PD^nP^{-1} \text{(Daddel, Genetics)}$$

- The columns of matrix P are formed by using the eigenvectors of matrix A. D is the diagonal matrix where the diagonal elements are the eigen values of A
- Suppose we know dominant allele "G" is associated with increased crop production and we want to determine the genetic proportion of the 10^{th} generation of crops (X_{10}), given that the initial genetic proportion of the crop population is:

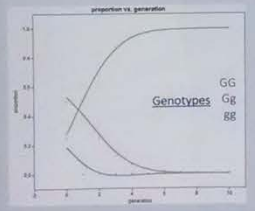
$$\frac{1}{4}GG, \frac{1}{2}Gg, and\ \frac{1}{4}gg.$$

$$\text{Then } X_0 = \begin{bmatrix} 1/4 \\ 1/2 \\ 1/4 \end{bmatrix} \text{ and } A = \begin{bmatrix} 1 & .5 & 0 \\ 0 & .5 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$A^{10} = PD^{10}P^{-1}$$

$$= \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1^{10} & & \\ & .5^{10} & \\ & & 0^{10} \end{bmatrix} * \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$



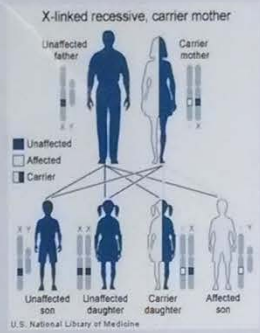
Now that we have A^{10}, we can solve $X_{10} = A^{10}X_0$

$$X_{10} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/2 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Thus 100% of the 10th generation crops will have the desired "G" gene.

X-Linked Inheritance

- 40% of the genes found on the X chromosome are associated with diseases (Thomson).
- X-linked inheritance differs from autosomal because the female determines two alleles and the male determines only one. A male receives only one X chromosome (XY) while a female receives two (XX).



- We chose to have at least 1 parent with the dominant gene in order to simulate the extinction of a recessive X-linked disease
- The genotypes in the left column are the offspring and the top row are the parenting phenotypes.

	XY XX	XY Xx	XY xx	xY XX	Xy Xx
XX	1/2	1/4	0	0	0
XX	0	1/4	1/2	1/2	1/4
XX	0	0	0	0	1/4
XX	1/2	1/4	0	1/2	1/4
XX	0	1/4	1/2	0	1/4

These proportions can be represented as a matrix:

$$A = \begin{bmatrix} 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/2 & 1/2 & 1/4 \\ 0 & 0 & 0 & 0 & 1/4 \\ 1/2 & 1/4 & 0 & 1/2 & 1/4 \\ 0 & 1/4 & 1/2 & 0 & 1/4 \end{bmatrix}$$

- Suppose the genotype proportion of an initial population is $\frac{1}{6}XX, \frac{1}{6}Xx, \frac{1}{6}xx, \frac{1}{4}xY, \frac{1}{4}Xy$
- Using $X_n = A^n X_0$, we solve for the future genotype proportions of the nth future generation:

$$X_n = \begin{bmatrix} 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/2 & 1/2 & 1/4 \\ 0 & 0 & 0 & 0 & 1/4 \\ 1/2 & 1/4 & 0 & 1/2 & 1/4 \\ 0 & 1/4 & 1/2 & 0 & 1/4 \end{bmatrix}^n \begin{bmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/4 \\ 1/4 \end{bmatrix}$$

$$\lim_{n \to \infty} X_n = \begin{bmatrix} 5/32 \\ 5/16 \\ 1/32 \\ 3/8 \\ 1/8 \end{bmatrix}$$

Notice as "n" approaches infinity, the genotypes converge to constant values.

Conclusions

- Markov chains prove to be an effective method in finding future genotype proportions for both autosomal and X –linked inheritance.
- The shortcut method by diagonalization, creates a faster method of predicting future phenotypes through autosomal inheritance.
- This research can be used to limit undesired genes, or increase beneficial genes in future generations of a given population
- Some ideas for future research would be to adjust our current X-linked matrix model so the diagonalization shortcut method can be applied; or to create matrices that could account for all possible combinations of genotype crossing.

References

- Daddel, Ali (2000). Genetics. [ONLINE] Available at: http://www.math.ucdavis.edu/~daddel/linear_algebra_appl/Applications/Genetics/genetics/genetics.html. [Last Accessed 2/27/13].
- Nussbaum, Robert L., Roderick R. McInnes, Huntington F. Willard, Ada Hamosh, and Margaret W. Thompson. Thompson & Thompson Genetics in Medicine.Philadelphia: Saunders/Elsevier, 2007. Print.
- Top Picture taken from Shutterstock, Inc.

Contact Info

- Jaron Arbet – Jarbet09@winona.edu
- Shaun Miller – smiller10@winona.edu