جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية التربية
قسم الرياضيات

بعض المشاكل في الإحصاء الجبري حول النماذج الفايلوجينية

رسالة ماجستير مقدمة
إلى مجلس كلية التربية – جامعة بابل كجزء من متطلبات درجة الماجستير علوم في الرياضيات

من قبل
نبيل جواد حسن ال بطي

بإشراف
أ.م.د. سعد عبد ماضي

2009

Republic of Iraq
Ministry of Higher Education
and Scientific Research
University of Babylon
College of Education
Department of Mathematics

Some Problems in Algebraic Statistics on Phylogenetic Models

This Thesis of Master Submitted to the Council of
College of Education, University of Babylon
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Mathematics

By
Nabeel Jawad Hassan
Al-Buti

Supervised By
Assist.Professor
Dr. Saad Abed Madhi

2009

بسم الله الرحمن الرحيم

وقل رب زدني علما

صدق الله العلي العظيم

سورة طه (الآية 114)

الخلاصة

الإحصاء الجبري هو دراسة التشكيلات الجبرية التي تنشأ في الإحصاء متعدد المتغيرات المتقطعة. هده التشكيلات تدعى النماذج الإحصائية الجبرية وان النقاط الحقيقية الغير سالبة على هذه التشكيلات التي تمثل توزيعات احتمالية على المتغيرات العشوائية التي تاخد عدد منتهي من الحالات. مثل النماذج الإحصائية / التشكيلات الجبرية تظهر بشكل متكرر في علوم الحياة و الوراثة.

هذه الرسالة تعلقت بدراسة بعض المشاكل في الإحصاء الجبري, حيث تم دراسة بعض نماذج النشوء النوعي من منظور الهندسة الجبرية.

Abstract

Algebraic statistics is the study of the algebraic varieties which arise in discrete multivariate statistics. These varieties are called algebraic statistical models and non-negative real points on these varieies represent probability distributions on random variables that take a finite number of states. Such statistical models /algebraic varieties make frequent appearance in the genetics and biological sciences.

This thesis is concerned with the study of some problems in algebraic statistics, where some phylogenetic models are studied from perspective of algebraic geometry.

The goal of this thesis , we study some phylogenetic models and translated them into segre varieties and found their ideals by computing minors of the joint distribution.

The goal of this thesis , we study some phylogenetic models and translated them into segre varieties and found their ideals by computing minors of the joint distribution.

Abstract

Acknowledgements

I would like to express my thanks and gratitude to my supervisor Dr. Saad Abed Madh for his thoroughly guidance, for the valuable suggestions and for the great effort that he offered not only during preparation of the thesis, but also during the course.

Special thanks are due to the college of Education and the Department of Mathematics, especially the head of the Department Dr. Ifticnar M.T and Dr. Eman Samir for their support and encouragement during my postgraduate study.

I would like to thank Mr. Muhamed Bakir in the Department of Biology-college of science for women, University of Babylon for explaining some concepts in Biology.

I would like to thank Mr.Bahaa Muhsin in the Department of Computer science-college of science for women, University of Babylon for using Matlab, and Mr. Habeeb Abid and Mr.Hassan Kamil for their encouragement during the prepation of this the thesis.

I owe a large debt of gratitude to the dean of the college of education Dr. Mohamed Jassigm, the head of the Department Dr. Akram and other members of mathematics Department in Thi-Qar University for their excellent teaching, during my B.Sc degree and support during postgraduate study.

Finally, I am very grateful for my family for their care, interest and patience.

Nabeel Jawad Hassan

Contents

Examining Committee Certification

We certify that we have read this thesis entitled " On Algebraic Statistics " and , as an Examining Committee , we examined the student in its content, and what is related to it , and that in our opinion it is adequate with standing as a thesis for the degree of Master of science in Mathematics.

Chairman

Signature :
Name :
Title :
Date :

Member	**Member**
Signature :	Signature :
Name :	Name :
Title :	Title :
Date :	Date :

Member and Supervisor

Signature :
Name :
Title :
Date :

Approved by the Dean of the college of Education

Signature :
Name :
Title :
Date :

Supervisor Certification

I certify that this thesis was prepared under my supervision at the Department of Mathematics / College of Education / University of Babylon , as a partial fulfillment of the requirements for the degree of Master of science in Mathematics.

Signature :
Name : Dr. Saad Abed Madhi
Title : Assist. Professor
Date :

RECOMMENDATION OF THE HEAD OF THE DERARTMENT

In view of the available recommendation , I forward this research for debate by the Examining Committee.

Signature :
Name : Dr. Iftichar Mudhar Talb Al-Shara
Title : Assist. Professor
Date :

List of Symbols

K , Q'	**ring**
$\mathbf{C}^n$	**Complex n – dimensional vector space**
χ	**Algebraic set**
I	**Ideal**
α, β	**Vector**
$O_v(U)$	**Sheaf of k-algebras**
ψ	**Ring homomorphism**
Λ	**Matrix with size 2×2**
P	**Joint probability distribution**
p_{ij}, p_{ijk}	**Probability as states ij , ijk**
α_{ij}, β_k	**Probability distribution**
Θ , $\mathbf{R}^d$	**Parameter space**
$\phi, \phi_1, \phi_2 \phi_3$	**Parameterized map**
$\mathbf{P}^n$	**Projective space**
Δ_m	**Probability simplex**
[m]	**State space**
$\amalg$	**Independence**
D(P)	**Ring of polynomial in m inderminates**
M	**Statistical model**
P', P'', P'''	**Joint probability distribution matrix**
G	**Graph**
T	**Tree**
V(G)	**Vertex set**
E(G)	**Edge set**
σ	**Segre map**
ρ	**Subset of parameter space**
Ω	**Set of nucleotides**
R^m, D^{n+1}	**Vector spaces**
V	**Topological space**
Q	**Rational numbers**
Z	**Integer numbers**
Z'	**Set of the some integer numbers**
B	**Set of the some integer numbers**

Reference

[1] Alexandersson M. , Cawley S. and Pachter L. , " Species Gene Finding and Alignment with Generalized Pair Hidden Markov Model " Genome Research 13 , 496 – 502 , 2003 .

[2] Barnier W. and Feldman N. , "Introduction to Advanced Mathematics" prentice – Hall Inc , upper saddle River , New jersey 07458 , 2000 .

[3] Bocci C. , " Topics on Phylogenetic Algebriac geometry " , 2008 .

[4] Galliam J. A. , " Contemporary Abstract Algebra " , D. C. Heath and company , 1994 .

[5] Gathmann A. , " Algebraic Geometry " , Notes for a class , taught at the university of Kaiserslautern 2002 – 2003 .

[6] Harris J. , " Algebraic Geometry " , a first course Berlin , New York , springer – verlag , ISBN 978 – 0 – 387 – 97716 – 4 , 1997 .

[7] Lander et al E. S. , " Initial Sequencing and Analysis of the Human

Genome " , Nature 409 , 860 – 921 , 2001 .

[8] Le H.V. , "Introduction to Algebraic Geometry " , matematicky U stave AVCR.

[9] Lipschutz S. and Lipson M. , " Discrete Mathematics " , Mc Graw – Hall Companies Inc , 1997 . Shaums outlines.

[10] Karp R. M. , " Mathematical Challenges from Genomics and Molecular Biology " , Notes of American Mathematical Society 49 , 544 – 553 , 2002 .

[11] Milne J. S. , " Algebriac Geometry " , October 30 , 2003 . Notes. www.j**milne**.org/math/CourseNotes/math631

[12] Myers et al E. , " A whole – Genome Assembly of Drosophila " , Science 287 , 2196 – 2204 , 2000 .

[13] Ny J. L. , " Some Experiments with Algebraic Statistics " , May 23 , 2006. www.mit.edu/~jleny/docs/reports/**algebraicStats**.pdf

[14] Pachter L. and Sturmfels B. , " Mathematics of Phylogenetic " , UC Berkely , California University press , *SIAM Review* **49** (2007) 3-31.

References

[15] Ranested K. , Sturmfels B. and Sullivant S. , " Phylogenetic Algebriac Geometry " , arxiv : math / 0407033 VI [math . AG] , 2 Jul 2004 .

[16] Smith L. , " Linear Algebra " , Springer – Verlag New York,Inc.1998

[17] Sullivant S. , " Statistical Models are Algebraic Varieties " , Algebraic Statistical Short Course, 2007
www.math.harvard.edu/~seths/lecture1.**pdf** .

[18]. Sullivant S. , " Log – Linear Models , Toric Varieties and their Markov Bases ", Algebraic Statistical Short Course. 2007.
www.math.harvard.edu/~seths/lecture2.pdf

[19] Sullivant S. , " Toric Ideals in Algebraic Statistics " , Springer 2005. Ph.D Thesis. San Francisco State University

[20] West D. B. , " Introduction to Graph Theory " , Prantice – Hall Inc, upper saddle River , NJ 07458 , 2001 .

1.1 Algebraic Statistics

Algebraic statistics is concerned with the study of algebraic varieties which arise in discrete multivariate analysis . An algebraic variety is the zero set of a collection of polynomials . Discrete multivariate analysis is concerned with the statistical models for random variables that take finite number of states . The crucial starting point upon which most of algebraic statistics rests , is that most statistical models used in practice for analyzing discrete random variables are actually algebraic varieties . The hope of researchers in this area is that this translation of statistical problems into algebraic language will lead to new insights and suggest new , practical algebraic techniques for solving real – world data analysis problems . Besides merely being a point of application for algebraic techniques , these statistical problems also high light new directions of research and drive the need for better algorithms . thus , the algebraic statistics interplay seems to benfit both statistics and algebra . [19]

1.2 Literature Review

The term algebraic statistics was coined by Pistone et. al. , E.Riccomagno and H.Wynn in their book " Algebraic Statistics " press (2000) . They have computational algebra in statistics . Hosten(2003) deals with the ideals of adjacent minors . He gave a description , of the minimal primes of the ideal generated by the $2 \times 2-$adjacent minors of a generic matrix . Sullivant (2004) talked about toric ideals of phylogenetic invariants . He discussed statistical models of evolution as algebraic varieties in the space of joint probability distribution on the leaf colorations of a phylogenetic tree. Garcia (2004) researched algebraic geometry of Bayesian networks . He computed the ideal of all polynomial functions which vanish on the space of observable distributions implied by any Bayesian network on three observable variables and one hidden variable . Pachter and Sturmfels (2005) introduced the more recent book in the algebraic statistics , its called " Algebraic Statistics and computational Biology " . It become the umbrella term for statistical research involving algebraic techniques . Allman (2005) studied the idetifiability of tree topology for phylogenetic models , including covarion and mixture models . She established tree identifiability for a number of phylogenetic models , including a covarion model and a variety of mixture models with limited number classes . Wynn (2006) considered the algebraic methods in the design of experiments . Drton (2007) studied algebraic statistical models and showed that the computational algebraic geometry can be used to solve problems arising in statistical inference in algebraic models . Pachter and sturmfels (2007) studied the mathematics of phylogenomics . They connected between the biological concepts and mathematical concepts. They used statistics , probability , combinartorics and algebraic geometry to discuss specific problems and developments arising from phylogenomics . Pachter and Sturmfels (2008) took about tropical geometry of statistical models and discussed how the solutions to various inference problems depend on the model parameters .

1.3 Outline of the Thesis

This work is divided into four chapters . In chapter two , we introduced preliminary definitions on algebra , algebraic geometry , statistical models , algebraic statistical model , biological concepts and biological mathematics concepts with some examples . In chapter three we studied the phylogenetics algebraic geometry , the models of evolutionary , polynomials maps drived from a tree , some phylogenetic models and some familiar varieties and interpretation algebraic geometry for this models . Finally we introduced three phylogenetic models and studied them from perspective algebraic geometry. In this thesis , we study some phylogenetic models and translated them into segre varieties and found their ideals by computing minors of the joint distribution.

2.1 Algebraic concepts [8]

Algebraic geometry deals with spaces and varieties over arbitrary ring but the classical algebraic geometry deals mostly with a closed field K . For the simplicity we shall take K = C . The main philosophy is to associate appropriate geometric notions (points , sets , topology , mapping ,etc) with corresponding algebraic notions (ideals , rings , zariski topology , morphisms ,etc) and conversely , appropriate algebraic notions with corresponding geometry notions .

Definition (2.1.1) [2]

A ring $(K,+,\bullet,0)$ consists of a non – empty set K , binary operations $+$ and $\bullet$ on K , and a distinguished element 0 in K such that the following axioms hold for all x, y and z in K :

1. $x+y=y+x$
2. $(x+y)+z=x+(y+z)$
3. $(x\bullet y)\bullet z=x\bullet(y\bullet z)$
4. $x\bullet(y+z)=(x\bullet y+x\bullet z)$ and $(y+z)\bullet x=(y\bullet x+z\bullet x)$
5. $x+0=x$
6. For every element $x\in K$, there exists an element $w\in K$ such that $x+w=0$.

Definition (2.1.2) [4]

A subring I of a ring K is called a (two – sided) ideal of K if for every $r \in K$ and every $a \in I$ both ra and ar in I .

. Definition (2.1.3) [8]

Let C^n be the complex n-dimensional vector space. A set χ subset of C^n is called algebraic , if there exists a subset T subset of $C[Z_1, Z_2, ..., Z_n]$ such that χ is the zero set of T :

$$\chi = Z(T),$$

i.e. for any f in T and any $(Z_1, Z_2, ..., Z_n)$ in χ we have

$$f(Z_1, Z_2, ..., Z_n) = 0 .$$

Definition (2.1.4) [8]

The zariski topology on C^n is defined by specifying the closed sets in C^n (equivalently the open sets in C^n). The closed sets in C^n in this topology are precisely algebraic sets. Equivalently a set is said to be open in zariski topology, if it is a complement of an algebraic set.

Definition (2.1.5) [8]

An algebraic set χ is called irreducible , if it can not be represented as the union of two algebraic sets such that each of them is a proper subset in S .

Definition (2.1.6) [11]

Monomials are ordered by lexicographic (dictionary) order. Let $\alpha=(a_1,a_2,...,a_n)$, $\beta=(b_1,b_2,...,b_n)$ in N^n then : $\alpha>\beta$ and $X^\alpha>X^\beta$ if , in the vector difference $\alpha-\beta$ in Z^n, the left most non – zero entry is positive .

Lemma (2.1.7) [11]

Let S be subset of N^n, then the ideal I generated by $\{x^\alpha:\alpha\in S\}$ is the monomial ideal corresponding to $A\overset{def}{=}\{\beta\in N^n:\beta-\alpha\in N^n$, some $\alpha\in S\}$. In other words, a monomial is in I if and only if it is divisible by one of the X^α, $\alpha\in S$.

Definition (2.1.8) [11]

Let $f=a_{\alpha_0}X^{\alpha_0}+a_{\alpha_1}X^{\alpha_1}+...+...$

and $\alpha_0>\alpha_1>...$, $a_{\alpha_0}\neq 0$, then we define:

1. The multidegree of f to be multdeg(f) = α_0;
2. The leading coefficient of f to be LC(f) = a_{α_0} ;
3. The leading monomial of f to be LM(f) = X^{α_0} ;
4. The leading term of f to be LT(f) = $a_{\alpha_0}X^{\alpha_0}$.

Theorem (2.1.9) [11]

Every ideal is finitely generated .

Theorem (2.1.10) [11]

Every ideal has a standard basis , and it generates the ideal , if $\{g_1, g_2, ..., g_s\}$ is a standard basis for an ideal I , then f in I iff the remainder on division by the g_i is zero .

Definition (2.1.11) [11]

An ideal I is monomial if $\sum c_\alpha X^\alpha \in I$ and $c_\alpha \neq 0$ then $X^\alpha \in I$.

Definition (2.1.12) [11]

For non- zero ideal I in $K(X_1, ..., X_n)$ we let (LT(I)) be the ideal generated by { LT(f) : f $\in$I } .

Definition (2.1.13) [6]

Let F be a collection of polynomials in the ring $K(X)$. The variety or zero set of F is the set

$$V_F(K) = \{a \in K^m : f(a) = 0 \text{ , for all } f \in F\} .$$

Definition (2.1.14) [8]

An affine algebraic variety is an irreducible closed algebraic set with the induced topology .

Definition (2.1.15) [11]

Let V be a topological space and D a field . Suppose that for every open subset U of V we have a set $O_v(U)$ of function U → D then O_v is called a sheaf of D-algebras if it satisfies the following conditions :

1. for each c in D the constant function c is in $O_v(U)$ and if f , g in $O_v(U)$ then so also do $f \pm g$ and fg .
2. if U' is an open subset of U and $f \in O_v(U)$ then $f / U' \in O_v(U')$.
3. Let $U = \bigcup U_\alpha$ be an open covering of an open subset U of V , then a function f : U→D is in $O_v(U)$ if $f / U_\alpha \in O_v(U_\alpha)$ for all U_α in some open cover of U.

Definition (2.1.16) [11]

A pair (V, O_v) consisting of a topological space V and a sheaf of D – algebras will be called ringed space .

Definition (2.1.17) [4]

A ring homomorphism ψ from a ring K to a ring Q' is mapping from K to Q' that preserves the two ring operations that is :

$$\psi(a + b) = \psi(a) + \psi(b)$$

$$\psi(a \cdot b) = \psi(a) \cdot \psi(b) \text{ , for all } a, b \in K \text{ .}$$

Remark (2.1.18) [4]

A ring homomorphism that is both one – to – one and onto is called a ring isomorphism .

Definition (2.1.19) [8]

A rational map is a morphism which is only defined on some open subset of a variety .

Definition (2.1.20) [16]

Let A be an n × n matrix . The square matrix obtained from A by deleting the ith – row and jth – column of A is called the minor of the element $a_{i,j}$ of A , or the (i , j) – minor of the matrix A and is denoted by $\Lambda_{i,j}$.

Definition (2.1.21) [16]

Let A be a square matrix of size n . Then the determinant of A , denoted by $\det(A)$ is defined by the inductive formula :

$$\mathrm{Det}(A)=\sum_{j=1}^{n}(-1)^{1+j}a_{1,j}\det(\Lambda_{1,j}),\text{if A is a square matrix of size } n>1.$$

Definition (2.1.22) [5]

We defined projective n – space over D , denoted P^n , to be the set of all one – dimensional linear subspace of the vector space D^{n+1} .

Remark (2.1.23) [5]

Obviously , a one – dimensional linear subspace of D^{n+1} is uniquely determined by a non – zero vector in D^{n+1}. Conversely , two such vectors $a=(a_0,a_1,...,a_n)$ and $b=(b_0,b_1,...,b_n)$ in D^{n+1} span the same linear subspace if and only if they differ only by a common scalar i.e if $b = \lambda a$ for some non – zero λ in D . In other words ,

$P^n =\{(a_0,...,a_n):a_i \in D$, not all $a_i = 0\} / \sim$ with equivalence relation

$(a_0,...,a_n) \sim (b_0,...,b_n)$ if $a=\lambda b$ for some λ in D / {zero } and all i.

This is often written as :

$$P^n = (D^{n+1} \setminus \{zero\}) / (D \setminus \{zero\})$$

and the point q in P^n determined by $(a_0,...,a_n)$ is written as q = (a_0: … : a_n) . So the notation (a_0 : … : a_n) means that the a_i are not all zero , and that they are defined only up to a common scalar multiple. The a_i are called the homogenous coordinates of the point q.

Definition (2.1.24) [6]

The segre map may be defined as the map

$$\sigma : P^n \times P^m \longrightarrow P^{(n+1)(m+1)-1}$$

taking a pair of points ([X] , [Y]) $\in$ $P^n \times P^m$, to their product

$$\sigma : ([X_0 , X_1 , \dots , X_n] , [Y_0 , Y_1 , \dots , Y_m]) \rightarrow [X_0 Y_0 , X_0 Y_1 , \dots , X_i Y_j , \dots , X_n Y_m] .$$

Here P^n are projective vector spaces over some arbitrary field and the notation $[X_0 : X_1 : \dots : X_n]$ is that of homogenous coordinates on the space . That image of the map is a variety , called a segre variety .

Example (2.1.25) [6]

Let m = n = 1 , we get an embedding of the product of the projective line with itself in P^n . The image is a quadric , and is easily seen to contain two one – parameter families of lines . Over the complex numbers this is a quite general non – singular quadric .Letting $[Z_0 : Z_1 : Z_2 : Z_3]$, be the homogenous coordinates on P^n , this quadric is given as the zero locus of the quadratic polynomial given by the determinant

$$\det \begin{bmatrix} Z_0 & Z_1 \\ Z_2 & Z_3 \end{bmatrix} = Z_0 Z_3 - Z_1 Z_2 .$$

Definition (2.1.26) [11]

A polynomial $f(X_0 , X_1 , \dots , X_n)$ is said to be homogenous of degree d if it is a sum of terms $a_{i_0,i_1,\dots,i_n} X^{i_0} \dots X^{i_n}$, with $i_0 + i_1 + \dots + i_n = d$ equivalently , $f(tX_0,\dots,tX_n) = t^d f(X_0,\dots,X_n)$ for all $t \in D$, write $D[X_0,\dots,X_n]_d$ for the subspace of $D[X_0,\dots,X_n]$ of polynomial of degree d, then

$$D[X_0,\dots,X_n] = \oplus_{d\geq 0} D[X_0,\dots,X_n]_d \ ;$$

that is , each polynomial F can be written uniquely as a sum $F=\sum F_d$ with homogenous of degree d .

Definition (2.1.27) [11]

An ideal $I\subset D[X_0,...,X_n]$ is said to be homogenous if it contains with any polynomial f all the homogenous components of F, i.e. $f\in I$ then $f_d\in I$, for all d , such an ideal I is generated by homogenous polynomials .

2.2 Statistical Models [17]

In this thesis , all random variables under consideration with be discrete and have a finite state space . If the cardinality of the state space of random variable X is m , we may assume that the underlying state space is the set of first m nonnegative integers $[m] = \{1,2,\dots,m\}$. When speaking in generalities about random variables and statistical models for them , we will make this assumption. On the other hand , in partical settings X will often be a random variable vector $X = (X_1, X_2, \dots, X_n)$ with finite state space $[m_1] \times [m_2] \times \dots \times [m_n]$, where m_i is the cardinality of the state space of the random variable X_i . In other situations it will be more natural to use the set $\{0,1,2,\dots, m - 1\}$ as the underlying state space , or use a different finite set altogether , such as $\{A, C, G, T\}^n$ when discussing DNA sequence . Since X is a discrete random variable with a finite state space , the probability distribution function of X is a point in an m – dimensional vector space . It is the point

$$P = (\text{prob}(X = 1), \text{prob}(X = 2), \dots, \text{prob}(X = m)) \in R^m.$$

The coordinates of P satisfy the constraints $p_i \geq 0$, for all $i \in [m]$ and $\sum_{i=1}^{m} p_i = 1$ since P is a probability distribution .

Conversely , any point $P \in R^m$ that is satisfies $p_i \geq 0$ for all $i \in [m]$ and $\sum_{i=1}^{m} p_i = 1$ is the probability distribution of some random variable X which has state space $[m]$.

Definition (2.2.1) [17]

The probability simplex Δ_m consists of all possible probability distributions for a random variable with state space $[m]$. In symbols

$$\Delta_m = \{P \in R^m : p_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{m} p_i = 1\}$$

Definition (2.2.2) [17]

A statistical model M for a random variable with state space $[m]$ is a subset of the probability simplex

$$M \subseteq \Delta_m$$

Example (Binomial Random Variable) (2.2.3) [17]

Consider a random variable X with $m+1$ states and space $\{0,1,2,...,m\}$,such that

$$\text{Prob}(X = i) = \binom{m}{i}\theta^i (1-\theta)^{m-i}$$

where $\theta \in [0,1]$, is a parameter . The interpretation is that θ is the probability of getting a head in one flip of a biased coin and X counts the number of heads that appear in m flips of a biased coin. The model

M consists of all probability distribution $P \in \Delta_{m+1}$ that arise in this way for some choice of parameter $\theta \in [0,1]$. The model M for this Bernoulli random variable is a curve in the m – dimensional probability simplex Δ_{m+1} .

Example(Independence of Two Random Variables) (2.2.4) [17]

Suppose that $X = (X_1, X_2)$ is two dimensional random vector , with state space

$$[m_1] \times [m_2] = \{1,2,3,\ldots,m_1\} \times \{1,2,3,\ldots,m_2\}$$

The random variables X_1 and X_2 are independence if the joint distribution satisfies

$$\Pr ob(X_1 = i, X_2 = j) = \Pr ob(X_1 = i) \Pr ob(X_2 = j) \ldots\ldots\ldots.(2.1)$$

for all $i \in [m_1]$ and $j \in [m_2]$. The notation $X_1 \amalg X_2$ is often used to indicate that X_1 is independent of X_2 . The constraints on the joint probability distribution from equation (2.1) become more algebraic as we translate them to constraints on the probability coordinates . In particular a distribution $P \in \Delta_m$ is the distribution of independent random variables if and only if

$$p_{ij} = (\sum_{j=1}^{m_1} p_{ij})(\sum_{i=1}^{m_2} p_{ij}) \ldots\ldots\ldots\ldots(2.2)$$

for all $i \in [m_i]$ and $j \in [m_2]$.

The model of independent $M_{X_1 \amalg X_2} \subset \Delta_{m_1 m_2}$, for two random variables consists of all joint distributions satisfying equation (2.1) or (2.2) .

Remark (2.2.5) [17]

The Bernoulli model is presented in a parametric form (as the image of a polynomial mapping) , whereas the independence model is presented in an implicit form (via constraints on the probability distributions) .

2.3 Algebraic Varieties [17]

The symbols $p_1, p_2, ..., p_m$ denote indeterminates (also called polynomial variable or just variables) . To avoid confusion with random variables , we will always call these polynomial variables in – determinates . A monomial in the indeterminates $p_1, p_2, ..., p_n$ is expression of the form :

$$p_1^{u_1} . p_2^{u_2} ... p_m^{u_m} ,$$

where $u_1, u_2, ..., u_m$ are non – negative integers . Often we will use the notation P^u to denote this monomial where $u = (u_1, u_2, ..., u_m)$. A polynomial is an expression of the form

$$f = \sum_{u \in U} c_u P^u$$

where the index set U is finite and each coefficient c_u belongs to a field K . In our setting , the underlying field will usually be the real number R , though for computational purpose we will often need to work with the field of rational number Q . The set of all polynomials in indeterminates $p_1, p_2, ..., p_m$ with coefficients in the field D is denote by $D(p_1, p_2, ..., p_m)$ though often we will use the short hand $D(P)$ to denote this set of polynomials. The set $D(P)$ has the algebraic structure of a ring, because we can add and multiply polynomials, and these operations are well-behaved with respect to one another foe instance, multiplication of polynomials distributes over addition. The ring D[P] is called the ring of polynomials in m indeterminates.

Definition (2.3.1) [17]

Let F be a collection of polynomials in the ring $D(P)$. The variety or zero set of F is the set :

$V_D(F) = \{a \in D^m : f(a) = 0,\ \text{for all}\ f \in F\}$.

Remark (2.3.2) [17]

From the stand point of proving theorems about varieties , it is often necessary to work in the setting where $D = C$ the field of complex numbers . In this setting , the expression $V(F)$ denotes the complex variety $V_C(F)$.

Example (2.3.3) [17]

Let $g = p_1^2 + p_2^2 - 1$ and $f = p_1^2 - p_2$, then $V_R(f)$ is parabola in the plane and $V_R(g)$ is a circle in the plane . The varieties $V(f)$ and $V(g)$ consist of the complex points of a parabola and circle, respectively. The variety $V_R(f,g)$ consists of the real intersection points of the parabola $V_R(f)$ and the circle $V_R(g)$, of which there are two. The variety $V(f,g)$ consists of the complex intersection points of the parabola $V\ (f)$ and the circle $V\ (g)$, of which there are four .

Definition (2.3.4) [17]

Let $S \subset D^m$ be a subset of m – dimensional space. The ideal of S, denoted $I(S)$ consists of all polynomials in $D(P)$ that vanish on S.

$$I(S) = \{f \in K(P) : f(a) = 0 \text{ , for all } a \in S\} \text{ .}$$

Remark (2.3.5) [17]

The collection of polynomials I(S) has the algebraic structure of an ideal. That is, suppose that f,g $\in$I(S) and h $\in$ D[P] are arbitrary polynomials. Then we always have f+g $\in$I(S) and $h \cdot f \in I(S)$. A collection of polynomials I which statisfies these two conditions on its elements is called an ideal. Since the ideal I(S) consists of all polynomials that vanish on S, we must always have the containment

$$S \subseteq V(I(S))$$

Though the two sets need not be equal (or even have the same dimension).

Definition (2.3.6) [17]

The variety V(I(S)) is called the zariski closure of S. It is the smallest variety that contains S.

Theorem (2.3.7) [17]

For any ideal $I \subseteq D[P]$, there exist a finite collection of polynomials $F \subset I$ such that $I = \langle F \rangle$.

Remark (2.3.8) [17]

In our applications , it will almost always be to our advantage to find the entire ideal $I(S)$, rather than just a collection of polynomials F , such that $V(F) = V(I(S))$. Our algebraic perspective will always be to find the entire ideal $I(S)$ if at all possible , the motivation coming from the individual applications . For instance, in the case of the independence model $M_{X_1 \amalg X_2}$, it will be especially significant that the 2×2- minors generate the entire ideal of the model whereas the condition coming from the equation (2.2) don't .

2.4 Algebraic Statistical Models [17]

We can describe more general algebraic statistical models. In particular, statistical models are usually presented parametrically (the primary exception to this rule are the independence models) .

Definition (2.4.1) [17]

Let $\Theta \subset R^d$ be a parametric space and $\phi : \Theta \to \Delta_m$ a map. The image $M = \phi(\Theta) \subset \Delta_m$ is called a parametric statistical model.

Definition (2.4.2) [17]

A set $\theta \subseteq R^d$ is called a semialgebraic set , if there are two finite collections of polynomials $F \subset D(P)$ and $G \subset D(P)$ such that $\Theta = \{\theta \in R^d : f(\theta) = 0$, for all $f \in F$ and $g(\theta) \geq 0$, for all $g \in G\}$.

Remark (2.4.3) [17]

Given a collection $\phi_1, \phi_2, ..., \phi_m$ of m rational functions , we get a map $\phi : D^d \to D^m$ defined by

$$\phi(\theta) = (\phi_1(\theta), \phi_2(\theta), ..., \phi_m(\theta)) .$$

Whenever we speak of rational map ϕ , we mean a function from K^d to K^m which can be written in this form .

Definition (2.4.4) [17]

Let $\Theta \subset R^d$ be a semialgebraic set and $\phi : R^d \to R^m$ be a rational map such that $\phi(\Theta) \subseteq \Delta_m$. Then $M = \phi(\Theta)$ is a parametric algebraic statistical model .

Example (Binomial Random Variable) (2.4.5) [17]

Continuing the model for a Bernoulli random variable from example (2.2.3) we can identify this model as parametric algebraic statistical model . Here the semialgebraic set $\Theta \subset R$ is just the segment [0,1] . The components of the rational map ϕ are $\phi_i = \binom{m}{i}\theta^i(1-\theta)^{m-i}$.

Example(Independence of Two Random Variables)(2.4.6) [17]

Continuing our discussion of the model $M_{X_1 \amalg X_2}$ of two independence random variables from example (2.2.4) , we can identify this model as a parametric algebraic statistical model . This is easy to do using equation (2.2) , we treat the marginal distributions $\Pr ob(X_1)$ and $\Pr ob(X_2)$ as the free parameters that completely specify the joint distributions for X . Thus the parameter space Θ , consists of all vectors (α,β) such that $\alpha \in \Delta_{m_1}$ and $\beta \in \Delta_{m_2}$; that is , the parameter space is the product of simplices $\Theta = \Delta_{m_1} \times \Delta_{m_2}$ and the parameterizing map ϕ is given by

$$\phi_{ij}(\alpha,\beta)=\alpha_i\beta_j \ .$$

The image $\phi(\Theta)\subset\Delta_m$ consists of all independent distributions for X_1 and X_2.

Remark (2.4.6) [17]

Let $\phi:C^d\to C^m$ be a rational parameterization . It is a basic fact of algebraic geometry that the image set $\phi(C^d)$, while not , in general, a variety, is well – approximated by a variety .

Definition (2.4.7) [13]

An algebraic statistical model $f:R^d\to R^m$ is called a linear model if each of its coordinate polynomial $f_i(\theta)$ is linear function i.e there exist vectors $a_i\in R^m$ and real numbers b_i such that

$$f_i(\theta)=a_i^T\theta+b_i=\sum_{j=1}^{d}a_{ij}\theta_j+b_i \qquad \text{for each } i\in\{1,2,\ldots,d\} \ .$$

For this model , it is convenient to take the m linear functions $f_1(\theta),f_2(\theta),\ldots,f_m(\theta)$ such that their sum is the constants function 1 .

Definition (2.4.8) [18]

Let $A \in N^{d\times m}$ be a matrix of nonnegative integers and suppose that the vector (1,1,…,1) is in the row span of A . Let $h \in R^{m}_{>0}$ be a vector of positive real numbers. Let $\Theta = R^{d}_{>0}$ and let $\widehat{\phi}^{A,h}$ be the rational parameterization

$\widehat{\phi}^{A,h} : \Theta \to R^{m}$, $\phi_j^{A,h}(\theta) = (h_j \cdot \prod_{i=1}^{d} \theta^{a_{ij}}) \cdot Z(\theta)^{-1}$ such that

$Z(\theta) = \sum_{j=1}^{m} h_j \prod_{i=1}^{d} \theta_i^{a_{ij}}$, for each $\theta \in \Theta$ is the normalizing constant.

The toric model is the parametric algebraic statistical model , $M_{A,h} = \widehat{\phi}^{A,h}(\Theta)$ with our restrictions on the matrix A .

Example (Binomial Random Variable) (2.4.9) [18]

Let A be the $2 \times (m+1)$ matrix

$$A = \begin{pmatrix} 0 & 1 & 2 & . & . & . & m \\ m & m-1 & m-2 & . & . & . & 0 \end{pmatrix}$$

and let $h_i = \binom{m}{i}$, for $i = 0,1,2,...,m$. Then the model $M_{A,h}$ is the model of Bernoulli random variable , with m flips of a biased coin . Indeed, we have

$$\widehat{\phi}_j^{A,h} = Z(\theta)^{-1} \binom{m}{j} \theta_1^j \theta_2^{m-j} \; .$$

However , we know that

$Z(\theta) = \sum_{j=0}^{m} \binom{m}{j} \theta_1^j \theta_2^{m-j} = (\theta_1 + \theta_2)^m$, thus we can write

$$\widehat{\phi}_j^{A,h} = \binom{m}{j}\left(\frac{\theta_1}{\theta_1+\theta_2}\right)^j\left(\frac{\theta_2}{\theta_1+\theta_2}\right)^{m-j}$$

and substituting $\theta = \dfrac{\theta_1}{(\theta_1+\theta_2)}$, and $1-\theta = \dfrac{\theta_2}{\theta_1+\theta_2}$, yields the parameterization

$$\widehat{\phi}_j^{A,h} = \binom{m}{j}\theta^j(1-\theta)^{m-j}$$

Example(Independence of Two Random Variable)(2.4.10) [18]

Let A be the $(m_1+m_2)\times m_1m_2$ matrix with columns $(e_i,e_j)^T$ where in the first entry we take e_i to be a standard unite vector in R^{m_1} and in the second entry we take e_j to be a standard unite vector in R^{m_2}. For instance if $m_1=2$ and $m_2=4$, then

$$A=\begin{pmatrix}1&1&1&1&0&0&0&0\\0&0&0&0&1&1&1&1\\1&0&0&0&1&0&0&0\\0&1&0&0&0&1&0&0\\0&0&1&0&0&0&1&0\\0&0&0&1&0&0&0&1\end{pmatrix}$$

set $h=(1,1,...,1)$ the all ones vector . If we label the parameters corresponding to the first m_1 rows by $\alpha_1,\alpha_2,...,\alpha_{m_1}$ and the parameters corresponding to the last m_2 rows by $\beta_1, \beta_2, \ldots, \beta_{m_2}$ then this yields the parameterization

$$\phi_{ij}^{A}(\alpha,\beta) = Z(\alpha,\beta)^{-1} \cdot \alpha_i \cdot \beta_j$$

where $Z(\alpha,\beta) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \alpha_i \beta_j = (\sum_{i=1}^{m_1} \alpha_i) \cdot (\sum_{j=1}^{m_2} \beta_j)$, from which we deduce the expression

$$\phi_{ij}^{A}(\alpha,\beta) = \frac{\alpha_i}{\sum_{i=1}^{m_1} \alpha_i} \cdot \frac{\beta_j}{\sum_{j=1}^{m_2} \beta_j}$$, substituting $\hat{\alpha}_i = \frac{\alpha_i}{\sum_{i=1}^{m_1} \alpha_i}$,

$\hat{\beta} = \frac{\beta_j}{\sum_{j=1}^{m_2} \beta_j}$, yields the parameterization from section four with parameter set $R_{>0}^{m_1+m_2}$ replaced by $\Delta_{m_1} \times \Delta_{m_2}$.

Definition (2.4.11) [18]

Given A and h , the log – linear model consists of all probability distributions satisfying $\log p_i = h_i + \sum_{i=1}^{d} a_{ij} \theta_i$, with $\theta \in R^d$.

Remark (2.4.12) [18]

Let ϕ^A be the toric parameterizing map (ignoring h) which we consider as a map from $K^d \to K^m$. The Zariski closure of the image $\phi^A(K^d)$ is called an affine toric variety and denoted V_A . The vanishing ideal $I(\phi^A(K^d) \subset K(P)$ is called a toric ideal and is denoted I_A .

Proposition (2.4.13) [18]

The toric ideal I_A is generated by all monomial difference $P^u - P^v$, where $u, v \in N^m$ and $Au = Av$:

$I_A =< P^u - P^v : u, v \in N^m$, and $Au = Av$ > .

Example(Independence of Two Random Variable)(2.4.16) [18]

In the case of independent of two discrete random variables , the toric ideal I_A is generated by the 2×2 - minors of the point distribution matrix P_{ij} :

$$I_A =< p_{ij} p_{kl} - p_{il} p_{kj} : i, k \in [m_1] \text{ and } j, l \in [m_2] > .$$

2.5 Biological Concepts

The challenges in biology today are being shaped by powerful high – throughput technologies that have revealed the genomes of many organisms , global expression patterns of genes and detailed information about variation within populations . We are therefore able to ask , for the first time , fundamental questions about the evolution of genomes , the structure of genes and their regulation , and the connections between genotypes and phenotypes of individuals . The answers to these questions are all predicated on progress in a variety of computational , statistics , and mathematical fields . [10]

The Genome (2.5.1)

Every living organism has a genome , made up of deoxyribonucleic acids (DNA) arranged in a double helix , which encodes (in a way to be made precise) the fundamental ingredients of life . Organisms are divided into two major classes : eukaryotes (organisms whose cells contain a nucleus) and prokaryotes (for example bacteria) . In our discussion we focus on genomes of eukaryotes , and in particular the human genome . [7]

Eukaryotic genomes are divided into chromosomes . The human genome has two copies of each chromosome . There are 23 pairs of chromosomes , 22 outosomes (two copies each in both men and women) and two sex chromosomes which are denoted X and Y. Women have two X chromosomes , while men have one X and one

Y chromosome . Parents pass on a mosaic of their pair of chromosomes to their children . The sequence of (DNA) molecules in a genome is typically represented as a sequence of letters partitioned into chromosomes , from the four letter alphabet $\Omega=\{A,C,G,T\}$. These letters corresponding to the bases in double helix , that is the nucleotides (Adenine , Cytosine , Guanine and Thymine) . Since every base is paired with an opposite base (A with T , and C with G in the other half of the double helix) , in order to describe a genome it suffices to list the bases in only one strand . However , it is important to note that the two strands have a directionality which is indicated by the numbers $5'$ and $3'$ on the ends (corresponding to carbon atoms in the helix backbone) . The convention is to represent (DNA) in the $5' \to 3'$ direction . The human genome consists of approximately 2.8 billion bases , and has been long . Sequence assemble algorithms are then used to piece together these fragments . [12]

Despite the tendency to abstract genomes as strings over the alphabet Ω, one must not forget that they are highly structured : for example , certain subsequences within a genome corresponding to genes . These subsequences play the important role of encoding protein . Proteins are polymers made of twenty different types of amino acids . Within a gene triplets of (DNA) known as codons . Table (1) shows the 64 possible codens , and the twenty amino acids they code for . Each amino acids is represented by a three letter identifier (Phe = phenylalanine , Leu = Leucin …) . The three codons (TAA , TAG and

TGA) are special : instead of coding for an amino acid , they are used to indicate that the protein ends .

	T	C	A	G	
T	TTT→Phe TTC→Phe TTA→Lue TTG→Lue	TCT→Ser TCC→Ser TCA→Ser TCG→Ser	TAT→Tyr TAC→Tyr TAA→Stop TAG→Stop	TGT→Cys TGC→Cys TGA→Stop TGG→Trp	T
C	CTT→Leu CTC→Leu CTA→Leu CTG→Leu	CCT→Pro CCC→Pro CCA→Pro CCG→Pro	CAT→His CAC→His CAA→Gln CAG→Gln	CGT→Arg CGC→Arg CGA→Arg CGG→Arg	C
A	ATT→Ile ATC→Ile ATA→Ile ATG→Met	ACT→Thr ACC→Thr ACA→Thr ACG→Thr	AAT→Asn AAC→Asn AAA→Lys AAG→Lys	AGT→Ser AGC→Ser AGA→Arg AGG→Arg	A
G	GTT→Val GTC→Val GTA→Val GTG→Val	GCT→Ala GCC→Ala GCA→Ala GCG→Ala	GAT→Asp GAC→Asp GAA→Glu GAG→Glu	GGT→Gly GGC→Gly GGA→Gly GGG→Gly	G

Table (1) : The standard genetic code

In order to make protein , (DNA) is first copied into a similar molecule called messenger (RNA) (abbreviated mRNA) in a process called transcription . It is the (RNA) that is translated into protein . The entire process is referred to as expression . Protein can be structural elements or perform complex tasks (such as regulation of expression) by interacting with the many molecules and complexes in cells . Thus , the genome is a blueprint for life . An understanding of the genes , the function of their proteins , and their expression patterns is fundamental to biology . The human genome contains approximately (25.000) genes , although the exact number has still not been determined . While there are experimental methods for validating and discovering genes , there is still no known high throughput technology for accurately identifying all the genes in a genome . One the main difficulties lies in the fact that only a small protein of any genome is genic . For instance , less than 5% of the human genome is known to be functional . [1]

2.6 Phylogenetic concepts

Codons (2.6.1) [14]

Because of the genetic code , the set Ω^3 of all three – letter words over the alphabet $\Omega = \{A, C, G, T\}$ plays a special role in molecular biology . These words are called codons , with each triplet coding for one of 20 amino acids (Table (1)) . The map from 64 codons to 20 amino acids is not injective , and so multiple codons code for the same amino acid . Such codons are called synonymous . Eight amino acids have the property that the synonymous codons that code for them all agree in the first two positions . The third positions of such codons are called four – fold degenerate . The translation of a series of codons in a gene (typically a few hundred) results in three – dimensional folded protein .

Definition (2.6.2) [14]

A model for codons is statistical model whose state space is the 64 – element set Ω^3 . Selecting a model means specifying a family of probability distributions $P = (p_{ijk})$ on Ω^3 . Each probability distribution P is a $4 \times 4 \times 4$ - table of non – negative real numbers which sum to one . Geometrically , a distribution on codons is point P in the 63 – dimensional probability simplex

$\Delta_{63} = \{P \in R^{\Omega^3} : \sum_{ijk \in \Omega^3} p_{ijk} = 1 \text{ and } p_{ijk} \geq 0 \text{ for all } ijk \in \Omega^3\}$.

A model for codons is hence nothing but a subset M of the simplex Δ_{63}.

Remark (2.6.3) [14]

Statistically meaningful models are usually given in parametric form . If the number of parameters is d , then there is a set $\rho \subset R^d$ of allowed parameters and the model M is the image of a map ϕ from ρ into Δ_{63}.

Remark (2.6.4) [14]

Consider a (DNA) sequence of length $3m$ which has been grouped into m consecutive codons . Let u_{ijk} denote the number of occurrences of a particular codon ijk . Then our data is the $4 \times 4 \times 4$ -table $u = (u_{ijk})$.

Remark (2.6.5) [14]

Let M be the statistical model which stipulates that , for the sequence under consideration , the first two positions in a codon are independent from the third position .

Remark (2.6.6) [14]

Many of the amino acids are uniquely specified by the first two positions in any codon which represents that particular amino acid (see Table 1) .Therefore , third positions in synonymous codons tend to be independent of the first two .

Independence Model for Codons (2.6.7) [14]

Our independence model M has 18 free parameters . The set of allowed parameters is an 18 – dimensional convex polytope, namely , it is product

$$\rho = \Delta_{15} \times \Delta_3$$

Here Δ_{15} is the 15 – dimensional simplex consisting of probability distributions $\alpha = (\alpha_{ij})$ on Ω^2, and Δ_3 is the tetrahedron consisting of probability distributions $\beta = (\beta_k)$ on Ω . Our model M is parameterized by the map

$$\phi : \rho \to \Delta_{63} \ , \ \phi((\alpha,\beta))_{ijk} = \alpha_{ij} \cdot \beta_k \ .$$

Hence $M = \text{image}\ (\phi)$ is an 18 – dimensional algebraic subset inside the 63 – dimensional simplex . To test whether a given $4 \times 4 \times 4$- table P lies in M, we write that table as a two - dimensional matrix with 16 rows and 4 columns

$$P' = \begin{pmatrix} p_{AAA} & p_{AAC} & p_{AAG} & p_{AAT} \\ p_{ACA} & p_{ACC} & p_{ACG} & p_{ACT} \\ p_{AGA} & p_{AGC} & p_{AGG} & p_{AGT} \\ p_{ATA} & p_{ATC} & p_{ATG} & p_{ATT} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ p_{TTA} & p_{TTC} & p_{TTG} & p_{TTT} \end{pmatrix} .$$

Linear algebra furnishes the following characterizations of our model :

Proposition (2.6.8) [14]

For a point $P \in \Delta_{63}$, the following conditions are equivalent :

1. The distribution P lies in the model M .
2. The 16×4 matrix P' has rank one .
3. All 2×2 - minors of the matrix P' are zero .
4. $p_{ijk} \cdot p_{lmn} = p_{ijn} \cdot p_{lmk}$, for all nucleotides i, j, k, l, m, n .

Remark (2.6.9) [14]

In the language of algebraic geometry , the model M is known as the Segre variety . More precisely, M is the set of non – negative real points on the Segre embedding of $P^{15} \times P^{3}$ in P^{63}.

2.7 Evolutionary Trees

Since graphs play an important role in phylogenetics , we start recalling some basic facts about it .

Definition (2.7.1) [20]

A directed graph or digraph G is a triple consisting of a vertex set $V(G)$ an edge set $E(G)$, and a function assigning each edge an ordered pair of vertices . The first vertex of the ordered pair is the tail of the edge , and the second is the head , together , they are endpoints .

Remark (2.7.2) [3]

All graphs we consider, will have a finite set of vertices . If $e = \{u, \upsilon\}$ is an edge of graph G, then u and υ are adjacent and e is said to be incident with u and υ . The vertices u and υ are the ends of e . Let υ be a vertex of a graph G. The valency of $\upsilon, \nu(\upsilon)$, is the number of edges in G that are incident with υ . A path in a graph G is a sequence of distinct vertices $\upsilon_1, \upsilon_2, ..., \upsilon_k$ such that , for all $i = 1, ..., k-1$, υ_1 and υ_{i+1} are adjacent . If , in addition υ_1 and υ_k are adjacent , then the subgraph of G whose vertex set $\{\upsilon_1, \upsilon_2, ..., \upsilon_k\}$ and whose edge set is $\{(\upsilon_k, \upsilon_1)\} \cup \{(\upsilon_i, \upsilon_{i+1}) : i = 1, ..., k-1\}$ is a cycle . A graph is connected if each pair of vertices in G can be joined by a path : otherwise G is disconnected .

Definition (2.7.3) [3]

A tree $T=(V,E)$ is a connected graph with no cycles . A tree is a path graph if all vertices have valency at most two .

Theorem (2.7.4) [3]

Let $G=(V,E)$ be a graph . Then the following are equivalent

1. G is a tree ;
2. for any two vertices υ and u in V there exists a unique path in G from υ to u;
3. G is connected and $|V|=|E|+1$.

Definition (2.7.5) [9]

A binary tree T is defined as a finite set of elements , called nodes such that

1. T is empty (called the null tree or empty tree) , or
2. T contains a distinguished node R, called the root of T , and the remaining nodes of T form an ordered pair of disjoint binary trees T_1 and T_2 .

Remark (2.7.6) [3]

A vertex of valency one called a leaf . The set of leaves denote by L and defined $\tilde{V}:=V\setminus L$ the set of interior vertex . Similarly we denote by $\tilde{E}$ the set of interior edges . Two distinct leaves of a tree are said to form a cherry if they are adjacent to common vertex .

Definition (2.7.7) [3]

A rooted tree is that has exactly one distinguished vertex called the root which we denote by the letter r . For a rooted tree T we can define a natural partial order $\leq_T$ on the vertex set V by $\upsilon_i \leq_T \upsilon_j$ if the path from the root of T to υ_i includes υ_j .

Definition (2.7.8) [3]

An $X-$Tree τ is an ordered pair (T,ϕ), where T is a tree with vertex set V and $\phi : X \rightarrow V$ is a map with the property that , for each $\upsilon \in V \setminus \{r\}$ of valency at most two , $\upsilon \in \phi(X)$. An X – tree is also called a semi – labelled tree (on X) A phylogenetic tree is an X – tree (T,ϕ) with the property that ϕ is a bijection from X into the set of leaves of T . If , in addition , every interior vertex of T has degree three , τ is binary phylogenetic tree .

Remark (2.7.9) [3]

Unrooted phylogenetic trees are also biologically relevant because they are typically what the tree reconstruction methods generate .We just observe that it always possible to pass from an unrooted tree to a rooted one and viceversa . In particular , passing from the unrooted to rooted tree means to choose an internal vertex as root or add another vertex inside an edge and choose it as root .

Remark (2.7.10) [3]

In general , as X , we will use the set $\{1,2,...,n\}$, where each number will correspond to a specific species .

3.1 Evolutionary Models [15]

The basic object in phylogenetic model is a tree T which is rooted and has n labeled leaves . Each node of the tree T is random variable with k possible states (usually k is taken to be 2 , for the binary states {0,1} or 4 for the nucleotides {A,C,G,T} . At the root the distribution of the states is given by $\pi = (\pi_1, \pi_2, ..., \pi_k)$. On each edge e of the tree there is $k \times k$ transition matrix Λ_e whose entries are indeterminates representing the probabilities of transition (away from the root) , between the states . The random variables at the leaves are observed . The random variables at the interior nodes are hidden . Let N be the total number of entries of the matrices Λ_e and the vector π. These entries are called model parameters . For instance , if T is binary tree with n leaves then T has $2n - 2$ edges , and hence $N = (2n-2)k^2 + k$. In particular , there will be many constraints on these parameters , usually expressible in terms of linear equations and inequalities , so the set of statistically meaningful parameters is a polyhedron ρ in R^N . Sometimes these constraints are given by non – linear polynomials , in which case ρ would be a semi – algebraic subset of R^N . Specifying this subset ρ means choosing a model of evolution . Fix a tree T with n leaves . At each leaf we can observe k possible states , so there are k^n possible joint observations we can make at the leaves . The probability ϕ_σ of making a particular observation σ is a polynomials ϕ_σ . This map is denoted

$$\phi : R^N \to R^{k^n} .$$

The map ϕ depends only on the tree T and the number k. In real-word applications, the coordinates ϕ_σ represent probabilities , so they should be non – negative and sum to 1 . In other words , the rules of probability require that $\phi(\rho)$, lie in the standard $(k^n - 1)$ - simplex in R^{k^n} . In phylogenetic algebraic geometry we temporarily abandon this requirement . We keep things simpler and closer to the familiar setting of complex algebraic geometry , by replacing ρ and $\phi(\rho)$ by their Zariski closures in C^N and C^{k^n} respectively . The polynomials ϕ_σ are often homogenous and $\phi(\rho)$ is best regarded as sub variety of a projective space .

3.2 Polynomials Maps Derived From a Tree [15]

In this section we explain the polynomial map ϕ associated to a tree T and an integer $k \geq 1$. To make things as concrete as possible, let $k = 2$ and T be the tree on $n = 3$ leaves picture below.

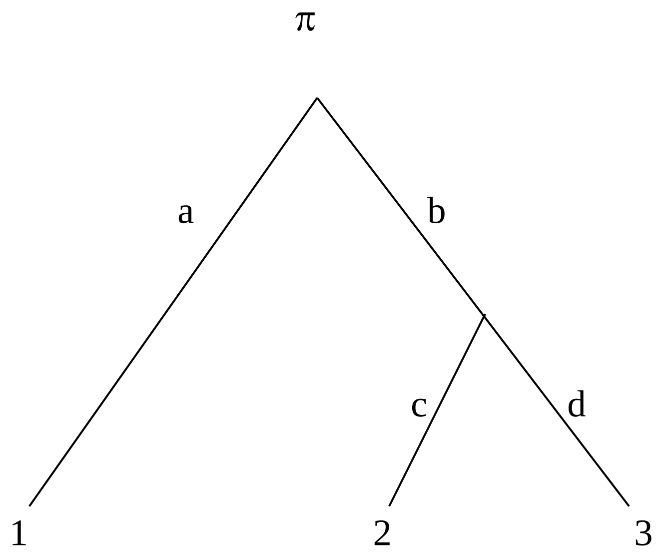


Figure (1)

The probability distribution at the root is unknown vector (π_0, π_1) . For each of the four edges of the tree , we have $2 \times 2-$ transition matrix :

$$\Lambda_a = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \;,\; \Lambda_b = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix}$$

$$\Lambda_c = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix} \ , \ \Lambda_d = \begin{pmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{pmatrix}$$

Altogether , we have introduced $N = 18$ parameters each of which represented a probability . But we regard them as unknown complex numbers . The unknown π_0 represents the probability of observing letter 0 at the root , and the unknown b_{01} represents the probability that the letter 0 gets changed to the letter 1 along the edge b . All transition are assumed to be independent events , so the monomial

$$\pi_u \cdot a_{ui} \cdot b_{uv} \cdot c_{vj} \cdot d_{vk}$$

represents the probability of observing the letter u at the root , the letter v at the interior node , the letter i at the leaf 1 , the letter j at the leaf 2 , and the letter k at the leaf 3 . Now , the probabilities at the root and the interior node are hidden random variable , while the probabilities at the three leaves are observed . This leads us to consider the polynomial

$$\phi_{ijk} = \pi_0 a_{0i} b_{00} c_{0j} d_{0k} + \pi_0 a_{0i} b_{01} c_{1j} d_{1k} + \pi_1 a_{1i} b_{10} c_{0j} d_{0k} + \pi_1 a_{1i} b_{11} c_{1j} d_{1k} .$$

This polynomial represents the probability of observing the letter i at the leaf 1 , the letter j at the leaf 2 , and the letter k at the leaf 3 . The eight polynomials ϕ_{ijk} specify our map

$$\phi : C^{18} \to C^8 \ .$$

In applications , where the parameters are really probabilities , one immediately replaces $C^{18}$by a subset ρ, for instance , the nine – dimensional cube in R^{18} defined by the constraints

$$\pi_0 + \pi_1 = 1 \ , \ \pi_0, \pi_1 \geq 0 \ ,$$

$$a_{00} + a_{01} = 1 \ , \ a_{00}, a_{01} \geq 0 \ , \qquad a_{10} + a_{11} = 1, \ a_{10}, a_{11} \geq 0$$

$$b_{00} + b_{01} = 1, \ b_{00}, b_{01} \geq 0 \ , \qquad b_{10} + b_{11} = 1, \ b_{10}, b_{11} \geq 0$$

$$c_{00} + c_{01} = 1 \ , \ c_{00}, c_{01} \geq 0 \ , \qquad c_{10} + c_{11} = 1, \ c_{10}, c_{11} \geq 0$$

$$d_{00} + d_{01} = 1, \ d_{00}, d_{01} \geq 0 \ , \qquad d_{10} + d_{11} = 1, \ d_{10}, d_{11} \geq 0 \ .$$

The polynomials ϕ_{ijk} are homogenous with respect to the different letters a, b, c, d and π . We can thus change our perspective and consider our map as a projective morphism

$$\phi : P^3 \times P^3 \times P^3 \times P^3 \times P^1 \to P^7 \ .$$

This morphism is subjective , and it is an instructive undertaking to examine its fibers . To underling the points made in the introduction , let us now cut down on the number of model parameters and replace the range of the morphisim by a natural subset ρ. For instance , let us define ρ by requiring that the four matrices are identical

$$\Lambda_a = \Lambda_b = \Lambda_c = \Lambda_d = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \ .$$

Equivalently , $\rho = P^3_{diag} \times P^1$, where P^3_{diag} is the diagonal of $P^3 \times P^3 \times P^3 \times P^3$. The restricted morphism $\phi \setminus \rho : P^3_{diag} \times P^1 \to P^7$ is given by the following eight polynomials :

$$\phi_{000} = \pi_0 a_{00}^4 + \pi_0 a_{00} a_{01} a_{10}^2 + \pi_1 a_{10}^2 a_{00}^2 + \pi_1 a_{10}^3 a_{11}$$

$$\phi_{001} = \pi_0 a_{00}^3 a_{01} + \pi_0 a_{00} a_{01} a_{10} a_{11} + \pi_1 a_{10}^2 a_{00} a_{01} + \pi_1 a_{10}^2 a_{11}^2$$

$$\phi_{010} = \pi_0 a_{00}^3 a_{01} + \pi_0 a_{00} a_{01} a_{10} a_{11} + \pi_1 a_{10}^2 a_{00} a_{01} + \pi_1 a_{10}^2 a_{11}^2$$

$$\phi_{011} = \pi_0 a_{00}^2 a_{01}^2 + \pi_0 a_{00} a_{01} a_{11}^2 + \pi_1 a_{10}^2 a_{01}^2 + \pi_1 a_{10} a_{11}^3$$

$$\phi_{100} = \pi_0 a_{00}^3 a_{01} + \pi_0 a_{01}^2 a_{10}^2 + \pi_1 a_{11} a_{10} a_{00}^2 + \pi_1 a_{10}^2 a_{11}^2$$

$$\phi_{101} = \pi_0 a_{00}^2 a_{01}^2 + \pi_0 a_{01}^2 a_{10} a_{11} + \pi_1 a_{11} a_{10} a_{00} a_{01} + \pi_1 a_{10} a_{11}^3$$

$$\phi_{110} = \pi_0 a_{00}^2 a_{01}^2 + \pi_0 a_{01}^2 a_{10} a_{11} + \pi_1 a_{11} a_{10} a_{00} a_{01} + \pi_1 a_{10} a_{11}^3$$

$$\phi_{111} = \pi_0 a_{01}^3 a_{00} + \pi_0 a_{01}^2 a_{11}^2 + \pi_1 a_{11} a_{10} a_{01}^2 + \pi_1 a_{11}^4 \, .$$

The image of $\phi \setminus \rho$ lies in the 5 – dimensional projective space subspace of P^7 defined by $\phi_{001} = \phi_{010}$ and $\phi_{101} = \phi_{110}$. **[15]**

Remark (3.2.1) [15]

Each coordinate of the map ϕ is given by a polynomial of degree equal to the number of edges of T plus one. If the root distribution is not parameter the degree of these polynomials is one less.

3.3 Some Models and Some Familiar Varieties [15]

Suppose that the number k of states , the number n of leaves and the tree T are fixed . The choice of model is then specified fixing a subset $\rho \subseteq C^N$. The set ρ comprises the allowed model parameters .

Definition (3.3.1) [15]

General Markov Model is the model $\rho = C^N$. All the transition matrices Λ_e are pairwise distinct , and there are no constraints on the k^2 entries of Λ_e .

Defiiniton (3.3.2) [15]

In Homogenous Model, the number of free parameters is small and independent of the tree .

Definition (3.3.3) [15]

Veronese Varieties appear as a special case of the homogenous model with no hidden nodes .

Definition (3.3.4) [15]

The secant varieties of a model amounts to taking the mixture of the model with itself .

Definition (3.3.5) [15]

Segre Varieties appear as a special case of the model with no hidden nodes .

Remark (3.3.6) [15]

Many of the evolutionary models are naturally embedding in-determinantal varieties , because the tree structure imposes rank constraints on matrices derived from probabilities observed at the leaves .

Example (3.3.7) [15]

We discuss an example which aims to demonstrate that phylogenetic trees arise when studying classical objects of algebraic geometry . Consider the Segre embedding of $P^1 \times P^1 \times P^1 \times P^1$ in P^{15} . This four – dimensional complex manifold is given by the familiar monomial parameterization

$$P_{ijkl} = u_i \cdot \upsilon_j \cdot \omega_k \cdot x_l \ , \qquad i, j, k, l \in [0,1]$$

Its prim ideal is generated by the $2 \times 2 -$ minors of the following three $4 \times 4 -$ matrices

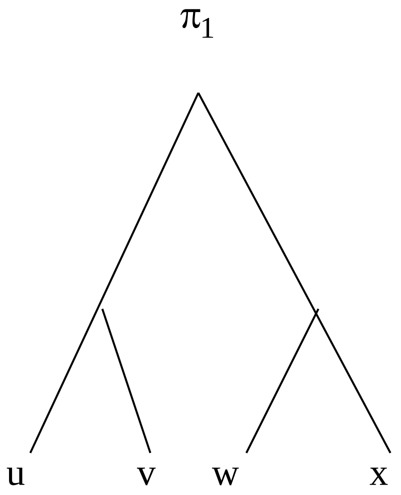
$$\begin{pmatrix} P_{0000} & P_{0001} & P_{0010} & P_{0011} \\ P_{0100} & P_{0101} & P_{0110} & P_{0111} \\ P_{1000} & P_{1001} & P_{1010} & P_{1011} \\ P_{1100} & P_{1101} & P_{1110} & P_{1111} \end{pmatrix}, \begin{pmatrix} P_{0000} & P_{0001} & P_{0100} & P_{0101} \\ P_{0010} & P_{0011} & P_{0110} & P_{0111} \\ P_{1000} & P_{1001} & P_{1100} & P_{1101} \\ P_{1010} & P_{1011} & P_{1110} & P_{1111} \end{pmatrix}$$

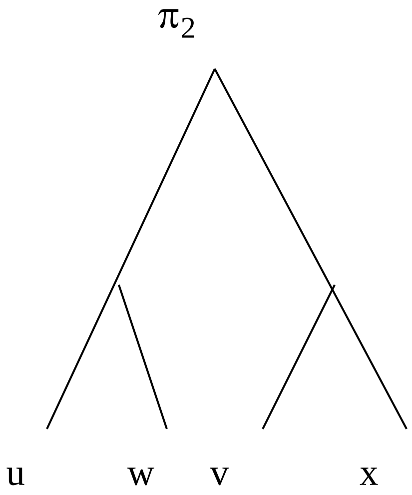
$$\begin{pmatrix} P_{0000} & P_{0010} & P_{0100} & P_{0110} \\ P_{0001} & P_{0011} & P_{0101} & P_{0111} \\ P_{1000} & P_{1010} & P_{1100} & P_{1110} \\ P_{1001} & P_{1011} & P_{1101} & P_{1111} \end{pmatrix}$$

These three matrices reflect the following three bracketings of the parameterization :

$$P_{ijkl} = ((u_i \cdot \upsilon_j) \cdot (\omega_k \cdot x_l)) = ((u_i \cdot \omega_k) \cdot (\upsilon_j \cdot x_l)) = ((u_i \cdot x_l) \cdot (\upsilon_j \cdot \omega_k))$$

And , of course , these three bracketings correspond to the three binary trees below .





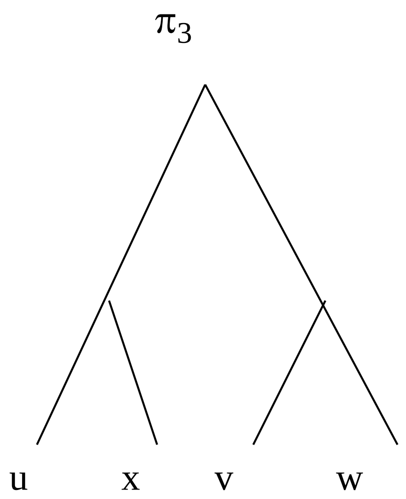


Figure (2)

Let χ denote the first secant variety of Segre variety $P^1 \times P^1 \times P^1 \times P^1$. Thus χ is the nine – dimensional irreducible of P^{15} consisting of all $2 \times 2 \times 2 \times 2 -$ tensors which have tensor rank at most 2 . The secant variety χ has the parametric representation $P_{ijkl} = \pi_0 \cdot u_{0i} \cdot \upsilon_{0j} \cdot \omega_{0k} \cdot x_{0l} + \pi_1 \cdot u_{1i} \cdot \upsilon_{1j} \cdot \omega_{1k} \cdot x_{1l}$.

This shows that the secant variety χ equals the general Markov model for the tree below .

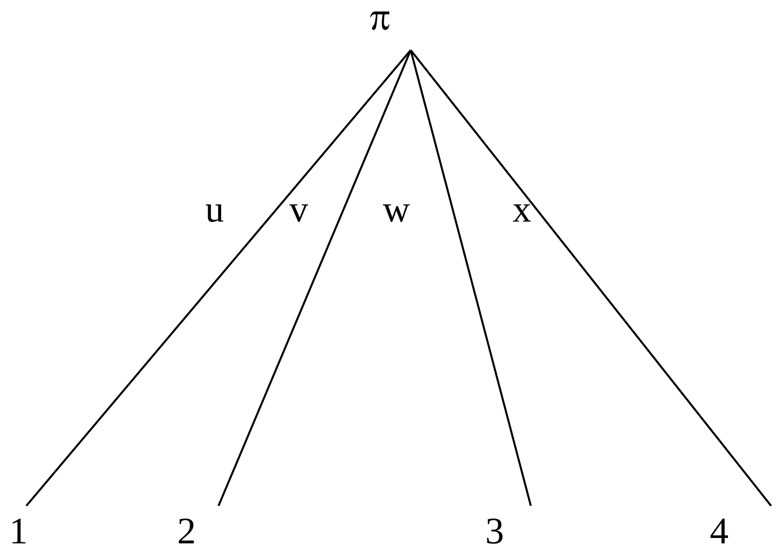


Figure (3)

The prim ideal of χ is generated by all the $3\times3-$ minors of the three matrices above . We write $\chi_{(12)}\chi_{(34)}$ for the variety defined by the $3\times3-$minors of the leftmost matrix , $\chi_{(13)}\chi_{(24)}$ for the variety of the $3\times3-$minors of the middle matrix , and $\chi_{(14)(23)}$ for the variety of the $3\times3-$minors of the rightmost matrix . Then we have , scheme – theoretically ,

$$\chi=\chi_{(12)(34)}\cap\chi_{(13)(24)}\cap\chi_{(14)(23)}$$

These three varieties are the general Markov model for the three binary trees depicted above . For instance , the determinantal variety $\chi_{(12)}\chi_{(34)}$ equals the general Markov model for the binary tree below .

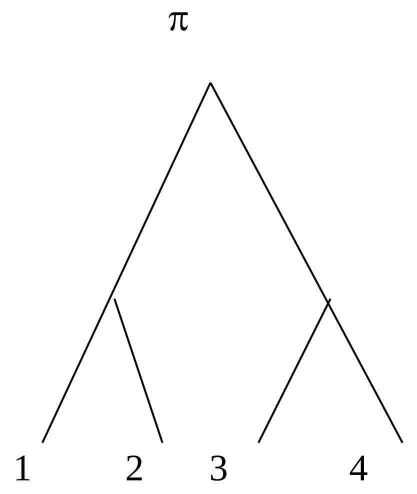


Figure (4)

Indeed , the standard parameterization ϕ of this model equals

$$P_{ijkl} = \pi_0 \cdot (a_{00}u_{0i}\upsilon_{0i} + a_{01}u_{1i}\upsilon_{1j}) \cdot (b_{00}\omega_{0k}x_{0l} + b_{01}\omega_{1k}x_{1l})$$
$$+ \pi_1 \cdot (a_{10}u_{1i}\upsilon_{1j} + a_{11}u_{1i}\upsilon_{1j}) \cdot (b_{10}\omega_{1k}x_{1l} + b_{11}\omega_{1k}x_{1l}) \ .$$

This representation shows that the leftmost $4 \times 4 -$matrix has rank at most 2 , and , conversely , every $4 \times 4 -$matrix of rank ≤ 2 can be written like this . We conclude that the general Markov model appears naturally when studying secant varieties of Segre varieties . It is instructive to redo the above calculations under the assumption

$u = \upsilon = \omega = x$. Then the ambient P^{15} gets replaced by the four – dimensional space P^4 with coordinates

$$\begin{aligned}
&P_0 = P_{0000}\\
&P_1 = P_{0001} = P_{0010} = P_{0100} = P_{1000}\\
&P_2 = P_{0011} = P_{0101} = P_{0110} = P_{1001} = P_{1010} = P_{1100}\\
&P_3 = P_{0111} = P_{1011} = P_{1101} = P_{1110}\\
&P_4 = P_{1111}
\end{aligned}$$

Under these substitutions , all three $4 \times 4 -$matrices reduce to the same $3 \times 3 -$matrix

$$\begin{pmatrix} P_0 & P_1 & P_2 \\ P_1 & P_2 & P_3 \\ P_2 & P_3 & P_4 \end{pmatrix}$$

The ideal of $2 \times 2 -$minors now defines the rational curve of degree four . This special Veronese variety is the small diagonal of the Segre variety $P^1 \times P^1 \times P^1 \times P^1 \subset P^{15}$. The secant variety of the rational normal curve is the cubic hypersurface in P^4 defined by the determinant of the $3 \times 3 -$matrix . Hence , the homogenous model satisfies

$$\chi = \chi_{(12)(34)} = \chi_{(13)(24)} = \chi_{(14)(23)}$$

In this chapter , we will introduce three phylogenetic models , and we will discuss these models using the language of algebraic geometry.

4.1 First Model

Suppose that we have a rooted tree with three leaves and no hidden interior nodes , on $\{A,C,G,T\}$ as given below.

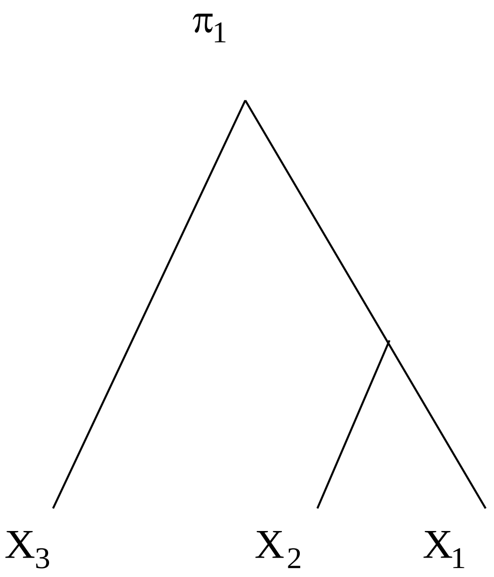


Figure (5)

Let (X_1, X_2, X_3) random vector with state space $\Omega^3 = \{A,C,G,T\}^3$, such that X_i, $i = 1,2,3$ with state space $\Omega = \{A,C,G,T\}$. Suppose that $\rho \subset R^d$, such that d is number of parameters ,then

$$\phi_1 : \rho \to \Delta_{63} \ , \ \phi_1(\rho) \subseteq \Delta_{63} \text{ is model .}$$

Let $X = (X_1, X_2)$, then X independent X_3 , such that X is a random vector with state space $\Omega^2 = \{A,C,G,T\}^2$, and X_3 random variable with state space $\Omega = \{A,C,G,T\}$.

Now

we have independence model $Y=(X,X_3)$, such that the probability simplex of X is Δ_{15} that consists of probability distribution $\alpha=\alpha_{ij}$ on Ω^2, and the probability simplex of X_3 is Δ_3 that consisting of the probability distribution $\beta=\beta_k$ on Ω . Then our model is parameterized by the map :

$$\phi_1=\rho\to\Delta_{63} \text{ , such that } \rho=\Delta_{15}\times\Delta_3$$

$$P_{ijk}=\phi_1((\alpha,\beta))_{ijk}=\alpha_{ij}\cdot\beta_k$$

Hence $M_1=\text{image }(\phi_1)$ is an 18 – dimensional algebraic subset inside the 63 – dimensional simplex .

Since Δ_4,Δ_{16} and Δ_{64} are subspaces of C^4,C^{16} and C^{64}, respectively , then

$$P^3=\Delta_4 \text{ , } P^{15}=\Delta_{16} \text{ and } P^{63}=\Delta_{64} \text{ .}$$

Therefore , our independence model M_1 become

$$\phi_1:P^{15}\times P^3\to P^{63}\text{, such that}$$

$\phi_1(P^{15}\times P^3)=M_1'$. Let $Z'=\{0,1,2,3\}=\{A,C,G,T\}$ and

$Z'\times Z'=\{A,C,G,T\}^2$ such that the symmetric elements are equals . Then our independence model M_1' becomes

$$\phi_1:P^{15}\times P^3\to P^{63}\text{ , such that}$$

$$\begin{gathered}\phi_1((\alpha_{00}:\alpha_{01}:\alpha_{02}:\alpha_{03}:\alpha_{10}:\alpha_{11}:\alpha_{12}:\alpha_{13}:\alpha_{20}:\alpha_{21}:\alpha_{22}:\alpha_{23}:\alpha_{30}:\\ \alpha_{31}:\alpha_{32}:\alpha_{33}),(\beta_0:\beta_1:\beta_2))=(\alpha_{00}\beta_0:\alpha_{00}\beta_1:\alpha_{00}\beta_2:\alpha_{01}\beta_0:\alpha_{01}\beta_1:\alpha_{01}\beta_2:\\ \alpha_{02}\beta_0:\alpha_{02}\beta_1:\alpha_{02}\beta_2:\alpha_{03}\beta_0:\alpha_{03}\beta_1:\alpha_{03}\beta_2:\alpha_{10}\beta_0:\alpha_{10}\beta_1:\alpha_{10}\beta_2:\alpha_{11}\beta_0:\alpha_{11}\beta_1:\\ \alpha_{11}\beta_2:\alpha_{12}\beta_0:\alpha_{12}\beta_1:\alpha_{12}\beta_2:\alpha_{13}\beta_0:\alpha_{13}\beta_1:\alpha_{13}\beta_2:\alpha_{20}\beta_0:\alpha_{20}\beta_1:\alpha_{20}\beta_2:\alpha_{21}\beta_0:\\ \alpha_{21}\beta_1:\alpha_{21}\beta_2:\alpha_{22}\beta_0:\alpha_{22}\beta_1:\alpha_{22}\beta_2:\alpha_{23}\beta_0:\alpha_{23}\beta_1:\alpha_{23}\beta_2:\alpha_{30}\beta_0:\alpha_{30}\beta_1:\alpha_{30}\beta_2:\\ \alpha_{31}\beta_0:\alpha_{31}\beta_1:\alpha_{31}\beta_2:\alpha_{32}\beta_0:\alpha_{32}\beta_1:\alpha_{32}\beta_2:\alpha_{33}\beta_0:\alpha_{33}\beta_1:\alpha_{33}\beta_2)\end{gathered}$$

Since $p_{ijk} = \alpha_{ij} \cdot \beta_k$, $i, j, k \in Z$, then

$$= (p_{000} : p_{001} : p_{002} : p_{003} : p_{010} : p_{011} : p_{012} : p_{013} : p_{020} : p_{021} : p_{022} : p_{023} : p_{030} : p_{031} : p_{032} : p_{033} : p_{100} : p_{101} : p_{102} : p_{103} : p_{110} : p_{111} : p_{112} : p_{113} : p_{120} : p_{121} : p_{122} : p_{123} : p_{130} : p_{131} : p_{132} : p_{133} : p_{200} : p_{201} : p_{202} : p_{203} : p_{210} : p_{211} : p_{212} : p_{213} : p_{220} : p_{221} : p_{222} : p_{223} : p_{230} : p_{231} : p_{232} : p_{233} : p_{300} : p_{301} : p_{302} : p_{303} : p_{310} : p_{311} : p_{312} : p_{313} : p_{320} : p_{321} : p_{322} : p_{323} : p_{330} : p_{331} : p_{332} : p_{333})$$

Transforming into matrix with size 8×8

$$P' = \begin{bmatrix} p_{000} & p_{001} & p_{002} & p_{003} & p_{010} & p_{011} & p_{012} & p_{013} \\ p_{020} & p_{021} & p_{022} & p_{023} & p_{030} & p_{031} & p_{032} & p_{033} \\ p_{100} & p_{101} & p_{102} & p_{103} & p_{110} & p_{111} & p_{112} & p_{113} \\ p_{120} & p_{121} & p_{122} & p_{123} & p_{130} & p_{131} & p_{132} & p_{133} \\ p_{200} & p_{201} & p_{202} & p_{203} & p_{210} & p_{211} & p_{212} & p_{213} \\ p_{220} & p_{221} & p_{222} & p_{223} & p_{230} & p_{231} & p_{232} & p_{233} \\ p_{300} & p_{301} & p_{302} & p_{303} & p_{310} & p_{311} & p_{312} & p_{313} \\ p_{320} & p_{321} & p_{322} & p_{323} & p_{330} & p_{331} & p_{332} & p_{333} \end{bmatrix}$$

Is the joint probability distribution matrix .

Since our model M_1', is independence model , therefore , the ideal of M_1', is generated by the $2 \times 2 -$minors of the joint probability distribution matrix P'. See program(A1) and program(A2) in appendix(A) .

In program(A1) , we find the positions matrix W with size 8×8,and for simplicity , we assume that the joint probability distribution matrix P'equal to the positions matrix W .In program(A2), we find the matrices with size 2×2 (minors) of P'.The minors of the joint probability distribution matrix P'

$$\begin{bmatrix} p_{312} & p_{313} \\ p_{332} & p_{333} \end{bmatrix}, \begin{bmatrix} p_{311} & p_{313} \\ p_{331} & p_{333} \end{bmatrix}, \begin{bmatrix} p_{311} & p_{312} \\ p_{331} & p_{332} \end{bmatrix} \cdots \begin{bmatrix} p_{300} & p_{301} \\ p_{320} & p_{321} \end{bmatrix}.$$

The ideal is

$$I = \langle p_{312}p_{333} - p_{313}p_{332}, p_{311}p_{333} - p_{313}p_{331}, p_{311}p_{332} - p_{312}p_{331}, \ldots, p_{300}p_{321} - p_{301}p_{320} \rangle$$

4.2 Second Model

Suppose that have a rooted tree with two leaves and no hidden interior nodes , on the set $\{A,C,G,T\}$. Let $X=(X_1,X_2)$ is random vector with state space $\Omega^2=\{A,C,G,T\}^2$, such that X_i , $i=1,2$ with state space $\Omega=\{A,C,G,T\}$, and $X_1 \amalg X_2$.

Now

$\Omega^2=\{A,C,G,T\}\times\{A,C,G,T\}=\{AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,$
$TC,TG,TT\}$

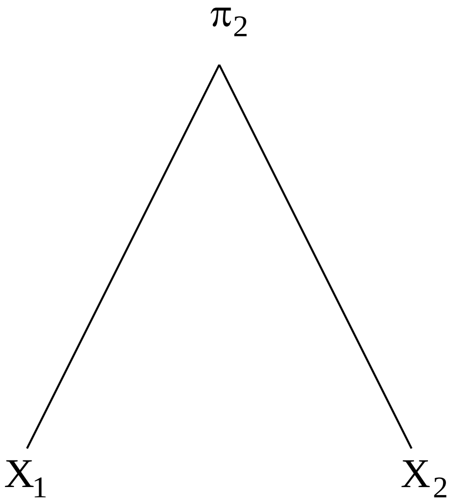


Figure (6)

Let $\rho\subset R^d$, where R^d is the parameter space and $\rho=\Delta_3\times\Delta_3$, $d=3+3=6$, is number of parameters , such that Δ_3 is the 3 – dimensional simplex consisting of probability distribution $\alpha=\alpha_i$ on Ω , and Δ_3 is the 3 – dimensional simplex consisting of probability distribution $\beta=\beta_j$ on Ω . Our model M_2 is parameterized by the map :

$\phi_2 : \rho \to \Delta_{15}$, such that $P_{ij} = \phi_2(\alpha,\beta)_{ij} = \alpha_i \cdot \beta_j$, $i, j \in \Omega$, with six parameters . Hence , $M_2 = \text{image}\ (\phi)$ is an 6 – dimensional algebraic subset inside the 15 – dimensional simplex . Let B = {0,1} is the binary set , and $B \times B = \{A, C, G, T\}$, such that $A = (0,0)$, $C = (0,1)$, $G = (1,0)$ and $T = (1,1)$. In the language algebraic geometry we have :

$P^3 = \Delta_4$, and $P^{15} = \Delta_{16}$, such that Δ_4 and Δ_{16} are subspace of the space C^4 and C^{16} , respectively . Then the parameterized map become

. $\phi_2 = P^3 \times P^3 \to P^{15}$ such that

$$\phi_2((\alpha_A : \alpha_C : \alpha_G : \alpha_T),(\beta_A : \beta_C : \beta_G : \beta_T)) = (\alpha_A\beta_A : \alpha_A\beta_C : \alpha_A\beta_G : \alpha_A\beta_T : \alpha_C\beta_A : \alpha_C\beta_C : \alpha_C\beta_G : \alpha_C\beta_T : \alpha_G\beta_A : \alpha_G\beta_C : \alpha_G\beta_G : \alpha_G\beta_T : \alpha_T\beta_A : \alpha_T\beta_C : \alpha_T\beta_G : \alpha_T\beta_T) = (\alpha_{00}\beta_{00} : \alpha_{00}\beta_{01} : \alpha_{00}\beta_{10} : \alpha_{00}\beta_{11} : \alpha_{01}\beta_{00} : \alpha_{01}\beta_{01} : \alpha_{01}\beta_{10} : \alpha_{01}\beta_{11} : \alpha_{10}\beta_{00} : \alpha_{10}\beta_{01} : \alpha_{10}\beta_{10} : \alpha_{10}\beta_{11} : \alpha_{11}\beta_{00} : \alpha_{11}\beta_{01} : \alpha_{11}\beta_{10} : \alpha_{11}\beta_{11})$$

Since $p_{ij} = \alpha_i \cdot \beta_j$, $i, j \in \Omega$

Then

$$= (p_{0000} : p_{0001} : p_{0010} : p_{0011} :{}_{0100} : p_{0101} : p_{0110} : p_{0111} : p_{1000} : p_{1001} : p_{1010} : p_{1011} : p_{1100} : p_{1101} : p_{1110} : p_{1111})$$

Transform into matrix with size 4×4 we have

$$P'' = \begin{bmatrix} p_{0000} & p_{0001} & p_{0010} & p_{0011} \\ p_{0100} & p_{0101} & p_{0110} & p_{0111} \\ p_{1000} & p_{1001} & p_{1010} & p_{1011} \\ p_{1100} & p_{1101} & p_{1110} & p_{1111} \end{bmatrix}$$ as the joint probability distribution

Since our model M_2'' , is independence model , therefore , the ideal of M_2'' , is generated by the $2\times 2-$minors of the joint probability distribution matrix P''. see program(B1) and program(B2) in appendix(B) .

In program(B1) , we found the positions matrix L with size 4×4,and for simplicity, we assume that the joint probability distribution matrix P''equal to the positions matrix L .In program(B2), we find the matrices with size 2×2 (minors) of P''.

The minors of the joint probability distribution matrix $P''$are

$$\begin{bmatrix} p_{1010} & p_{1011} \\ p_{1110} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1011} \\ p_{1101} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1010} \\ p_{1101} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1010} & p_{1011} \\ p_{1110} & p_{1111} \end{bmatrix},$$

$$\begin{bmatrix} p_{1000} & p_{1011} \\ p_{1100} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1010} \\ p_{1100} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1011} \\ p_{1101} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1011} \\ p_{1100} & p_{1111} \end{bmatrix},$$

$$\begin{bmatrix} p_{1000} & p_{1001} \\ p_{1100} & p_{1101} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1010} \\ p_{1101} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1010} \\ p_{1100} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1001} \\ p_{1100} & p_{1101} \end{bmatrix}$$

The ideal is

$$I=\langle p_{1010}p_{1111}-p_{1011}p_{1110}, p_{1001}p_{1111}-p_{1011}p_{1101}, p_{1001}p_{1110}-p_{1010}p_{1101}, p_{1000}p_{1111}-p_{1011}p_{1100}, p_{1000}p_{1110}-p_{1010}p_{1100}, p_{1000}p_{1101}-p_{1001}p_{1100}, p_{1000}p_{1110}-p_{1010}p_{1100}\rangle$$

4.3 Third Model

Suppose that we have a rooted tree with three leaves and not hidden interior nodes on a binary set $\{0,1\}$. Let $X=(X_1,X_2,X_3)$ is a random vector with state space $[m]^3=\{0,1\}^3$, such that X_i , $i=1,2,3$ with state space $[m_i]=\{0,1\}$. The state space of the random vector X is $[m]^3=\{0,1\}^3=\{000,001,010,011,100,101,110,111\}$

Now

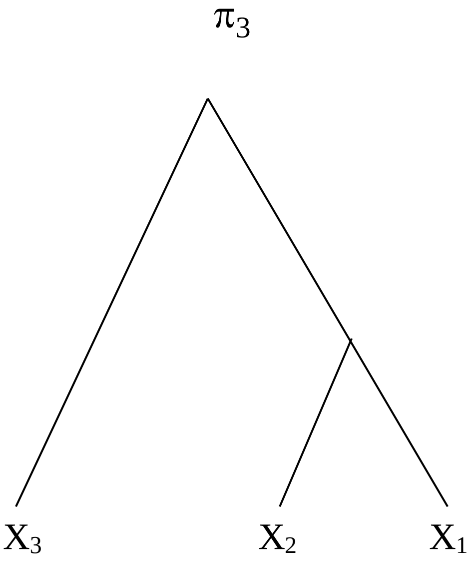


Figure (7)

Let $\rho\subset R^d$, such that $d=m_1+m_2+m_3=2+2+2=6$, then $\rho\subset R^6$

Let $Y=(X_1,X_2)$ which is independence of X_3 , i.e $Y\amalg X_3$, Y random vector with state space $[m]^2=\{0,1\}^2$, X_3 random variable with state space $[m]=\{0,1\}$. Let $\rho=\Delta_3\times\Delta_1$, such that Δ_3 is a $3-$dimensional probability simplex that consist of probability distribution $\alpha=\alpha_{ij}$ on $\{0,1\}^2$, and Δ_1 is $1-$dimensional probability simplex that consists of probability distribution $\beta = \beta_k$, on $\{0,1\}$. Our model M_3 is parameterized by the map :

$$\phi_3 : \rho \to \Delta_7 \text{ , such that}$$

$P_{ijk} = \phi_3(\alpha,\beta)_{ijk} = \alpha_{ij} \cdot \beta_k$, $i,j,k \in \{0,1\}$.

In the language algebraic geometry , we have :

$P^3 = \Delta_4$, $P^1 = \Delta_2$, and $P^7 = \Delta_8$, such that Δ_4 , Δ_2 and Δ_8 are subspaces of the space C^4 , C^2 , and C^8 , respectively . Then our model M_3 become

$$\phi_3 : P^3 \times P^1 \to P^7 \text{, such that}$$

$\phi_3((\alpha_{00} : \alpha_{01} : \alpha_{10} : \alpha_{11}),(\beta_0,\beta_1)) = (\alpha_{00}\beta_0 : \alpha_{00}\beta_1 : \alpha_{01}\beta_0 : \alpha_{01}\beta_1 : \alpha_{10}\beta_0 : \alpha_{10}\beta_1 : \alpha_{11}\beta_0 : \alpha_{11}\beta_1)$

. Since

$p_{ijk} = \alpha_{ij}\beta_k$. Then

$p_{ilk} = (p_{000} : p_{001} : p_{010} : p_{011} : p_{100} : p_{101} : p_{110} : p_{111})$.

Transform into matrix with size 2×4 we have

$$P''' = \begin{bmatrix} p_{000} & p_{001} & p_{010} & p_{011} \\ p_{100} & p_{101} & p_{110} & p_{111} \end{bmatrix}$$,as the joint probability distribution matrix.

Since our model M_3''' , is independence model , therefore , the ideal of M_3''' is generated by $2 \times 2 -$minors of joint probability distribution P''' .

The minors are

$$\begin{bmatrix} p_{000} & p_{001} \\ p_{100} & p_{101} \end{bmatrix}, \begin{bmatrix} p_{001} & p_{010} \\ p_{101} & p_{110} \end{bmatrix}, \begin{bmatrix} p_{000} & p_{011} \\ p_{100} & p_{111} \end{bmatrix}, \begin{bmatrix} p_{010} & p_{011} \\ p_{110} & p_{111} \end{bmatrix}, \begin{bmatrix} p_{000} & p_{010} \\ p_{100} & p_{110} \end{bmatrix},$$

$$\begin{bmatrix} p_{001} & p_{011} \\ p_{110} & p_{111} \end{bmatrix}$$, and hence the ideal is

$$I = \langle P_{000}P_{101} - P_{001}P_{100}, P_{001}P_{110} - P_{010}P_{101}, P_{000}P_{111} - P_{011}P_{100}, P_{010}P_{111} - P_{011}P_{110}, P_{001}P_{111} - P_{011}P_{110}, P_{000}P_{110} - P_{010}P_{100} \rangle$$

4.4 Conclusion

We found bijective between independence models and segre varieties, and found the ideals of those models(segre varieties).

4.5 Future Work

1. We will study graphical models from perspective algebraic geometry and commutative algebra, and will find the ideal of those models by grobner bases.

2. We will discuss conditional models from perspective algebraic geometry and commutative algebra, and will find the ideal of those models by grobner bases.

Appendix(A)

Program (A1)

```
clc
W=cell(8,8);
for i=1:8
for j=1:8
y1=num2str(i);
y2=strcat('a',y1);
y3=num2str(j);
y4=strcat(y2,y3);
W{i,j}=y4;
end
end
W
```

```
W =

'a11'    'a12'    'a13'    'a14'    'a15'    'a16'    'a17'    'a18'
'a21'    'a22'    'a23'    'a24'    'a25'    'a26'    'a27'    'a28'
'a31'    'a32'    'a33'    'a34'    'a35'    'a36'    'a37'    'a38'
'a41'    'a42'    'a43'    'a44'    'a45'    'a46'    'a47'    'a48'
'a51'    'a52'    'a53'    'a54'    'a55'    'a56'    'a57'    'a58'
'a61'    'a62'    'a63'    'a64'    'a65'    'a66'    'a67'    'a68'
'a71'    'a72'    'a73'    'a74'    'a75'    'a76'    'a77'    'a78'
'a81'    'a82'    'a83'    'a84'    'a85'    'a86'    'a87'    'a88'
```

For easy the program we assume that the joint probability distribution matrix P' equal to positions matrix as

$$P' = W = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} & a_{18} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} & a_{28} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & a_{37} & a_{38} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} & a_{48} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & a_{57} & a_{58} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & a_{67} & a_{68} \\ a_{71} & a_{72} & a_{73} & a_{74} & a_{75} & a_{76} & a_{77} & a_{78} \\ a_{81} & a_{82} & a_{83} & a_{84} & a_{85} & a_{86} & a_{87} & a_{88} \end{bmatrix} =$$

$$\begin{bmatrix} p_{000} & p_{001} & p_{002} & p_{003} & p_{010} & p_{011} & p_{012} & p_{013} \\ p_{020} & p_{021} & p_{022} & p_{023} & p_{030} & p_{031} & p_{032} & p_{033} \\ p_{100} & p_{101} & p_{102} & p_{103} & p_{110} & p_{111} & p_{112} & p_{113} \\ p_{120} & p_{121} & p_{122} & p_{123} & p_{130} & p_{131} & p_{132} & p_{133} \\ p_{200} & p_{201} & p_{202} & p_{203} & p_{210} & p_{211} & p_{212} & p_{213} \\ p_{220} & p_{221} & p_{222} & p_{223} & p_{230} & p_{231} & p_{232} & p_{233} \\ p_{300} & p_{301} & p_{302} & p_{303} & p_{310} & p_{311} & p_{312} & p_{313} \\ p_{320} & p_{321} & p_{322} & p_{323} & p_{330} & p_{331} & p_{332} & p_{333} \end{bmatrix}$$

Program (A2)

```
   function minors=h(a)
clc
kk=0;
 minors=zeros(2,40320);
for i1=1:8
  a1=m(a,i1);
    for i2=1:7
         a2=m(a1,i2);
for i3=1:6
     a3=m(a2,i3);
for i4=1:5
     a4=m(a3,i4);
for i5=1:4
     a5=m(a4,i5);
for i6=1:3
     a6=m(a5,i6)
    kk=kk+1
    pause
%     minors(1,kk)=a6(1,1);
%     minors(1,kk+1)=a6(1,2);
%     minors(2,kk)=a6(2,1);
```

```
%      minors(2,kk+1)=a6(2,2);
%      kk=kk+2;

end
end
end
end
    end
end
 save('d:\minorss.dat','minors','-ASCII');
end
%-----------------------------------------------
-----

function f=m(a,k)
[r,c]=size(a);
k1=1;
k2=0;
f=cell(r-1,c-1);
for i=1:r

    for j=1:c
        if (i~=1)&&(j~=k)
```

```
            k2=k2+1;
            if (k2>(c-1))
                k2=1;
                k1=k1+1;

            end
            f(k1,k2)=a(i,j);

        end
    end
end

end
```

$$\begin{bmatrix} a_{77} & a_{78} \\ a_{87} & a_{88} \end{bmatrix},\begin{bmatrix} a_{76} & a_{78} \\ a_{86} & a_{88} \end{bmatrix},\begin{bmatrix} a_{76} & a_{77} \\ a_{86} & a_{87} \end{bmatrix}\cdots\begin{bmatrix} a_{71} & a_{72} \\ a_{81} & a_{82} \end{bmatrix}.$$

These minors equal to

$$\begin{bmatrix} p_{312} & p_{313} \\ p_{332} & p_{333} \end{bmatrix},\begin{bmatrix} p_{311} & p_{313} \\ p_{331} & p_{333} \end{bmatrix},\begin{bmatrix} p_{311} & p_{312} \\ p_{331} & p_{332} \end{bmatrix}\cdots\begin{bmatrix} p_{300} & p_{301} \\ p_{320} & p_{321} \end{bmatrix}$$

respectively.

Appendix(B)

Program (B1)

```
clc
L=cell(4,4);
for i=1:4
for j=1:4
y1=num2str(i);
y2=strcat('a',y1);
y3=num2str(j);
y4=strcat(y2,y3);
L{i,j}=y4;
end
end
L
```

```
L =

    'a11'    'a12'    'a13'    'a14'
    'a21'    'a22'    'a23'    'a24'
```

```
   'a31'    'a32'    'a33'    'a34'
   'a41'    'a42'    'a43'    'a44'
```

For easy the program we write the joint probability distribution matrix P'' as following figure

$$P'' = L = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} p_{0000} & p_{0001} & p_{0010} & p_{0011} \\ p_{0100} & p_{0101} & p_{0110} & p_{0111} \\ p_{1000} & p_{1001} & p_{1010} & p_{1011} \\ p_{1100} & p_{1101} & p_{1110} & p_{1111} \end{bmatrix}$$

Program (B2)

```
function minors=h(a)
clc
kk=0;
 minors=zeros(2,24);
for i1=1:4
   a1=m(a,i1);
      for i2=1:3
             a2=m(a1,i2);
      kk=kk+1
      pause
%        minors(1,kk)=a6(1,1);
%        minors(1,kk+1)=a6(1,2);
%        minors(2,kk)=a6(2,1);
%        minors(2,kk+1)=a6(2,2);
%        kk=kk+2;

end
end
save('d:\minorss.dat','minors','-ASCII');
end
```

```
%-----------------------------------------------
-----


function f=m(a,k)
[r,c]=size(a);
k1=1;
k2=0;
f=cell(r-1,c-1);
for i=1:r

    for j=1:c
        if (i~=1)&&(j~=k)

            k2=k2+1;
            if (k2>(c-1))
                k2=1;
                k1=k1+1;

            end
            f(k1,k2)=a(i,j);

        end
```

$$\begin{bmatrix} a_{33} & a_{34} \\ a_{43} & a_{44} \end{bmatrix}, \begin{bmatrix} a_{32} & a_{34} \\ a_{42} & a_{44} \end{bmatrix}, \begin{bmatrix} a_{32} & a_{33} \\ a_{42} & a_{43} \end{bmatrix}, \begin{bmatrix} a_{33} & a_{34} \\ a_{43} & a_{44} \end{bmatrix},$$

$$\begin{bmatrix} a_{31} & a_{34} \\ a_{41} & a_{44} \end{bmatrix}, \begin{bmatrix} a_{31} & a_{33} \\ a_{41} & a_{43} \end{bmatrix}, \begin{bmatrix} a_{32} & a_{34} \\ a_{42} & a_{44} \end{bmatrix}, \begin{bmatrix} a_{31} & a_{34} \\ a_{41} & a_{44} \end{bmatrix},$$

$$\begin{bmatrix} a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix}, \begin{bmatrix} a_{32} & a_{33} \\ a_{42} & a_{43} \end{bmatrix}, \begin{bmatrix} a_{31} & a_{33} \\ a_{41} & a_{43} \end{bmatrix}, \begin{bmatrix} a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix}.$$

These minors equal

$$\begin{bmatrix} p_{1010} & p_{1011} \\ p_{1110} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1011} \\ p_{1101} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1010} \\ p_{1101} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1010} & p_{1011} \\ p_{1110} & p_{1111} \end{bmatrix},$$

$$\begin{bmatrix} p_{1000} & p_{1011} \\ p_{1100} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1010} \\ p_{1100} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1011} \\ p_{1101} & p_{1111} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1011} \\ p_{1100} & p_{1111} \end{bmatrix},$$

$$\begin{bmatrix} p_{1000} & p_{1001} \\ p_{1100} & p_{1101} \end{bmatrix}, \begin{bmatrix} p_{1001} & p_{1010} \\ p_{1101} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1010} \\ p_{1100} & p_{1110} \end{bmatrix}, \begin{bmatrix} p_{1000} & p_{1001} \\ p_{1100} & p_{1101} \end{bmatrix}$$

respectively.