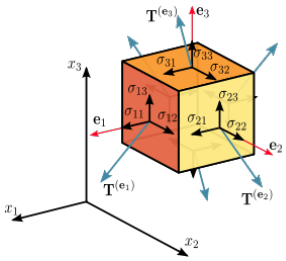
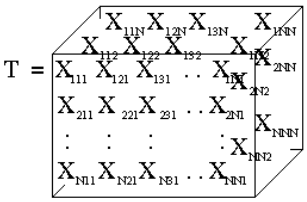
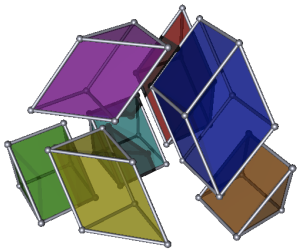
UNIVERSITÀ
DI SIENA
1240

Department of Information Engineering and Mathematics

University of Siena – Italy







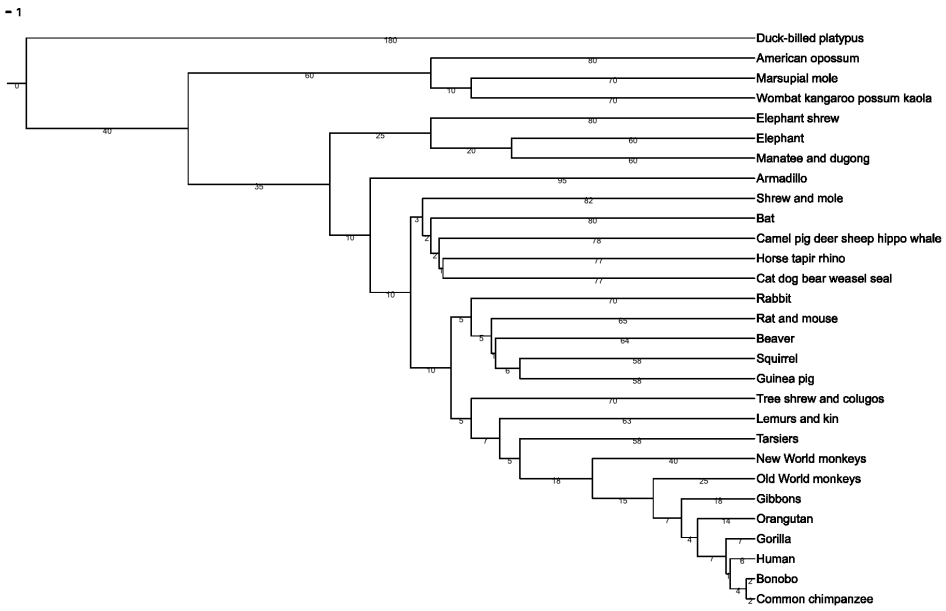
Cristiano Bocci

cristiano.bocci@unisi.it

Lectures Notes

on

Applications of Tensors



Winter School in Algebraic Geometry and Applications

February 17th-20th, 2014

Konkuk University, Seoul, Korea

Contents

Lecture 1. Basic facts in Algebraic Statistics

Statistics is the science of data analysis. What do statisticians do with their data? They build models of the process that generated the data and, in what is known as statistical inference, draw conclusions about this process.

In Algebraic Statistics the inference tools look different from those found in many texts on mathematical statistics or computational biology: they are written in the language of abstract algebra. The algebraic language for statistics clarifies many of the ideas central to the analysis of discrete data, and, within the context of biological sequence analysis, unifies the main ingredients of many widely used algorithms.

Algebraic Statistics is a new field, less than two decades old, whose precise scope is still emerging. The term itself was coined by Giovanni Pistone, Eva Riccomagno and Henry Wynn, with the title of their book [47]. That book explains how polynomial algebra arises in problems from experimental design and discrete probability, and it demonstrates how computational algebra techniques can be applied to statistics. This application reveals many important aspects. In particular if a statistical model can be described by a set of polynomials, then we can use Gröbner bases to study such set and then, as a consequence, we have the chance to use softwares for Computer Algebra as `Co.Co.A.`, `Macaulay2` and `Singular` to study such statistical models.

If [47] can be considered the european point of view in Algebraic Statistics, it is mandatory to notice that american researchers introduced new algebraic and geometric tools in the study of statistical models: polytopes, graph theory, tropical geometry, secant varieties. This use mainly concerned with applications on sequences alignment, statistical inference, Phylogenetics, Enumerative Biology (see, for example, [46]).

Year after year different aspects of statistical models show their counter-part in some algebraic and/or geometric discipline bringing new ideas to attack specific problems in Statistics or to describe statistical events.

Thus we can now talk about an Algebra/Geometry-Statistics Dictionary:

Statistics	Algebra
Independence	Segre Variety
Binomial Random Variable	Rational Normal Curve
Log-linear Model	Toric Variety
Mixture Model	Secant Variety
ML Estimation	Tropicalization
Design	Zero-dimensional Scheme
⋮	⋮

This dictionary is far to be complete and new correspondences can be added, as we will see, for the case of the concept of identifiability, in Lecture 3.

1.1 Statistical models

Definition 1.1.1. *A* **statistical model** *is a family of probability distributions on some state space.*

Our state space is finite and denoted by $[m] = \{1, 2, \dots, m\}$. A probability distribution on $[m]$ is a point of the probability simplex

$$\Delta_{m-1} := \{(p_1, \dots, p_m) \in \mathbb{R}^m : \sum_{i=1}^{m} p_i = 1, \quad p_i \geq 0 \quad \forall i\}.$$

We can use also the language of random variables: let X be a random variable with values in $[m]$, then the distribution of X is the point

$$\Big(\text{Prob}(X = 1), \text{Prob}(X = 2), \dots, \text{Prob}(X = m)\Big) \in \Delta_{m-1}.$$

Definition 1.1.2. *A* **statistical model** *is a subset $\mathcal{M}$ of Δ_{m-1}.*

Example 1.1.3 (Binomial random variable)**.** Let X be the random variable describing the number of heads in m flips of a biased coin (then the state space is $\{0, 1, \dots, m\}$). Let us denote the unknown bias by $\theta \in [0, 1]$. If, for example, the coin is a "legal" one, then $\theta = \frac{1}{2}$. If $\theta \geq \frac{1}{2}$ then head is favorite, while for $\theta \leq \frac{1}{2}$ tail will be favorite. Thus the probability to observe j heads in m flips of the coin is given by:

$$\text{Prob}(X = j) = \binom{m}{j} \theta^j (1 - \theta)^{m-j} \quad \theta \in [0, 1]$$

Hence the associated statistical model, denoted $\mathcal{M}_m$, is the set of vectors $\left((1-\theta)^m, \binom{m}{1}\theta(1-\theta)^{m-1}, \ldots, \theta^m\right)$ as θ varies in $[0,1]$, i.e.

$$\mathcal{M}_m = \left\{ \left((1-\theta)^m, \binom{m}{1}\theta(1-\theta)^{m-1}, \ldots, \theta^m\right) : \quad \theta \in [0,1] \right\}.$$

Definition 1.1.4. *Let $S \subseteq K[p_1, p_2, \ldots, p_m]$ be a set of polynomials in the p_i's and consider*

$$Z_\Delta(S) = Z_K(S) \cap \Delta_{m-1}$$

where $Z_K(S) = \{a \in K^m : f(a) = 0, \forall f \in S\}$ is the zero set of S. Then $Z_\Delta(S)$ is an **algebraic statistical model**.

Example 1.1.5. Consider the Hankel matrix

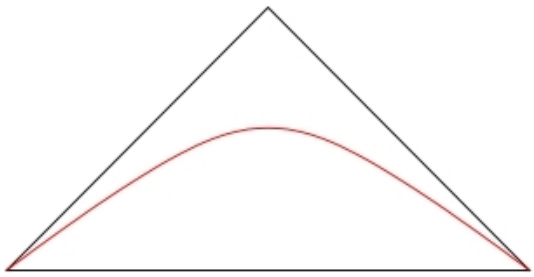
$$M_m = \begin{pmatrix} p_0 & \frac{p_1}{\binom{m}{1}} & \frac{p_2}{\binom{m}{2}} & \cdots & \frac{p_{m-1}}{\binom{m}{m-1}} \\ \frac{p_1}{\binom{m}{1}} & \frac{p_2}{\binom{m}{2}} & \frac{p_3}{\binom{m}{3}} & \cdots & \frac{p_{m-1}}{\binom{m}{m-1}} \end{pmatrix}$$

and let S_m be the set of 2×2 minors of M_m

$$S_m = \left\{ \frac{p_0 p_2}{\binom{m}{2}} - \frac{p_1^2}{\binom{m}{1}^2}, \frac{p_0 p_3}{\binom{m}{3}} - \frac{p_1^2}{\binom{m}{1}\binom{m}{2}}, \ldots \right\}.$$

Then the model $\mathcal{M}_m$ of Example 1.1.3 is exactly $Z_{\Delta_m}(S_m)$.

In the case $m = 2$ one has $\mathcal{M}_2 = Z_\Delta(4p_0p_2 - p_1^2)$ which can be represented in the following way



Definition 1.1.6. *Consider a continuos function*

$$\begin{array}{rccc} f: & \Theta \subseteq \mathbb{R}^d & \to & \mathbb{R}^m \\ & \theta = (\theta_1, \ldots, \theta_d) & \mapsto & \left(f_1(\theta), \ldots, f_m(\theta)\right) \end{array}$$

Then $f(\Theta) \subset \Delta_{m-1}$ is a **parametric statistical model**, *denoted by $\mathcal{M}_f$.*

Remark 1.1.7. *It is clear that the $i-th$ component of f, f_i, corresponds to the probability p_i of the event i.*

The unknowns $\theta_1, \ldots, \theta_d$ represent the **model parameters**. In most cases of interest, d is much smaller than m. The parameter vector $\theta_1, \ldots, \theta_d$ ranges over a suitable non-empty open subset $\Theta \in \mathbb{R}^d$ which is called the **parameter space of the model** f. We assume that the parameter space Θ satisfies the condition

$$f_i(\theta) > 0 \text{ for all } i \in [m] \text{ and } \theta \in \Theta \tag{1.1}$$

Under these hypotheses, the following two conditions are equivalent:

$$f(\Theta) \subseteq \Delta_{m-1} \iff \sum f_i(\theta) = 1. \tag{1.2}$$

If (1.2) holds, then our model is simply the set $f(\Theta)$. If it does not hold,we pass to consider the normalization $\frac{1}{\sum f_i(\theta)}(f_1(\theta), \ldots, f_m(\theta))$.

Definition 1.1.8. *Suppose that the components f_i's of the function f in Definition 1.1.6 are of the form $f_1 = \frac{g_1}{h_1}, \ldots, f_m = \frac{g_m}{h_m}$, where $g_i, h_i \in \mathbb{R}[\theta_1, \ldots, \theta_d]$ and $\Theta \subset \mathbb{R}^d$ is a semi-algebraic set. Then $f(\Theta) \subset \Delta_{m-1}$ is a* **parametric algebraic statistical model**.

In this model, after multiplying for the common denominator, each coordinate function p_i is a polynomial in the d unknowns, which means it has the form

$$f_i(\theta) = \sum_{a \in \mathbb{N}^d} c_a \theta_1^{a_1} \theta_2^{a_2} \cdots \theta_d^{a^d}$$

where all but finitely many of the coefficients $c_a \in \mathbb{R}$ are zero. Here (1.2) is an identity of polynomial functions, which means that all non-constant terms of the polynomials f_i cancel, and the constant terms add up to 1.

Example 1.1.9. The model $\mathcal{M}_m$ of Example 1.1.3 is a parametric algebraic statistical model since

$$\Theta = [0, 1]$$

and

$$f_j = p_j = \binom{m}{j} \theta^j (1 - \theta)^{m-j}.$$

Let us introduce now the so-called mixture and secant models. Consider two random variables X, H with $X \in [m]$, $H \in [r]$, such that

- $X_{|H=i}$ is represented by a distribution $\underline{p_i}$ in $\mathcal{M}_X$
- H is represented by a distribution $\pi \in \Delta_{r-1}$
- H is hidden, the probability of states at X is given by $\sum_{i=1}^{r} \pi_i \underline{p_i}$

Definition 1.1.10. *A* **mixture model** *is defined as*

$$Mixt^r(\mathcal{M}) = \left\{ \sum_{i=1}^{r} \pi_i \underline{p_i} \,|\, \pi \in \Delta_{r-1}, \underline{p_1}, \ldots, \underline{p_r} \in \mathcal{M} \right\}$$

The **secant model** *is defined as*

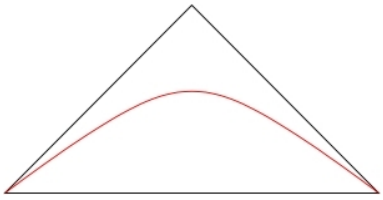
$$Sec^r(\mathcal{M}) = \overline{Mixt^r(\mathcal{M})} \cap \Delta_{m-1}$$

Example 1.1.11. Consider the model $\mathcal{M}_2$ of a binomial random variable on state $[0, 1, 2]$ as defined in Example 1.1.3. Then, as already said, $\mathcal{M}_2$ is a subset of Δ_2 given by the equation

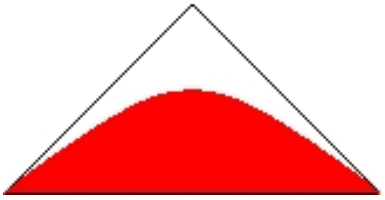
$$det \begin{pmatrix} p_0 & \frac{p_1}{2} \\ \frac{p_1}{2} & p_2 \end{pmatrix} = 0.$$

The second secant model of $\mathcal{M}_2$, $Sec^2(\mathcal{M}_2)$, is the subset of Δ_2 given by the condition

$$det \begin{pmatrix} p_0 & \frac{p_1}{2} \\ \frac{p_1}{2} & p_2 \end{pmatrix} \geq 0.$$



$\mathcal{M}_2$



$Sec^2(\mathcal{M}_2)$

1.2 Maximum likelihood estimation

Our data are tipically given in the form of a sequence of observations

$$i_1 i_2 \cdots i_N \qquad i_j \in [m]. \tag{1.3}$$

Each data point i_j is an element from our state space $[m]$. The integer N, which is the length of the sequence, is called the **sample size**. Assuming that the observations (1.3) are independent and identically distributed samples, we can summarize the data (1.3) in the data vector

$$u = (u_1, \ldots, u_m)$$

where u_k =number of indices $j \in N$ such that $i_j = k$. Hence u is a vector in $\mathbb{N}^m$ with $u_1 + u_2 + \cdots + u_m = N$.

Definition 1.2.1. *The* **empirical distribution** *is* $\frac{u}{N} = \frac{1}{N}(u_1, \ldots, u_m)$. *The entries* $\frac{u_i}{N}$*'s are the* **observed relative frequencies** *of the various otucomes.*

Remark 1.2.2. It is obvious that the empirical distribution is a point in the probability simplex Δ_{m-1}.

Let us fix our attention in the parametric statistical model $\mathcal{M}_f$. We say that this model is a **good fit** for the data u if there exists a parameter vector $\theta \in \Theta$ such that the probability distribution $f(\theta)$ is very close, in a statistical meaningful sense, to the empirical distribution. Suppose we draw N times at random (independently and with replacement) from the set $[m]$ with respect to the probability distribution $f(\theta)$. Then the probability of observing the sequence (1.3) equals

$$L(\theta) = f_{i_1}(\theta) f_{i_2}(\theta) \cdots f_{i_N}(\theta) = f_1(\theta)^{u_1} f_2(\theta)^{u_2} \cdots f_m(\theta)^{u_m}. \tag{1.4}$$

This expression depends on the parameter vector θ as well as the data vector u. However, we think of u as being fixed and then L is a function from Θ to the positive real numbers.

Definition 1.2.3. *$L(\theta)$ is the* **likelihood function** *for the model $\mathcal{M}_f$ for the data $i_1 i_2 \cdots i_N$.*

Remark 1.2.4. Note that any reordering of the sequence (1.3) leads to the same data vector u. Hence the probability of observing the data vector u is equal to

$$\frac{(u_1 + u_2 + \cdots + u_m)!}{u_1! u_2! \cdots u_m!} L(\theta) \tag{1.5}$$

which is called the **scaled likelihood function**.

Remark 1.2.5. The vector u plays the role of a **sufficient statistic** for the model $\mathcal{M}_f$. This means that the likelihood function $L(\theta)$ depends on the data (1.3) only through u.

In practice one often replaces the likelihood function by its logarithm

$$\ell(\theta) = \log(L(\theta)) = u_1 \cdot \log((f_1(\theta)) + u_2 \cdot \log((f_2(\theta)) + \cdots + u_m \cdot \log((f_m(\theta)) \tag{1.6}$$

This is the **log-likelihood function**. Note that $\ell(\theta)$ is a function from the parameter space $\theta \subseteq \mathbb{R}^d$ to the negative real numbers $\mathbb{R}_{<0}$.

The problem of maximum likelihood estimation is to maximize the likelihood function $L(\theta)$ in (1.4), or, equivalently, the scaled likelihood function (1.5), or, equivalently, the log-likelihood function $\ell(\theta)$ in (1.6). Here θ ranges over the parameter space $\theta \subseteq \mathbb{R}^d$. Formally, we consider the optimization problem:

$$\text{Maximize } l(\theta) \text{ subject to } \theta \in \Theta.$$

Definition 1.2.6. *A solution $\hat{\theta}$ of the optimization problem is a* **maximum likelihood estimate** *of θ with respect to the model $\mathcal{M}_f$ and data u.*

Sometimes, if the model satisfies certain properties, it may be that there always exists a unique maximum likelihood estimate $\hat{\theta}$. This happens for linear models and toric models, due to the concavity of their log-likelihood function, as we shall see in the Section 1.4. For most statistical models, however, the situation is not as simple. First, a maximum likelihood estimate need not exist (since we assume Θ to be open). Second, even if $\hat{\theta}$ exists, there can be more than one global maximum, in fact, there can be infinitely many of them. And, third, it may be very difficult to find any one of these global maxima. In that case, one may content oneself with a local maximum of the likelihood function. There are numerical methods for finding solutions to the maximum likelihood estimation problem such as the EM-algorithm.

1.3 Independence models and contingency tables

Let X and Y be discrete random variables. We assume that X takes on values in the set $[m]$ and Y takes on values in $[n]$. Their **joint probability distribution** is the $m \times n-$matrix $P = (p_{ij})$ where

$$p_{ij} = \text{Prob}(X = i \text{ and } Y = j).$$

The real numbers p_{ij}'s are probabilities, then they are non-negative and their sum is one. Thus P is a point in a probability simplex of dimension $mn - 1$, in symbols,

$$P \in \Delta_{mn-1} = \left\{ u \in \mathbb{R}^{m \times n} \ : \ \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij} = 1 \text{ and } u_{ij} \geq 0 \text{ for all } i, j \right\}.$$

The row sums of the matrix P give the distribution of the random variable X:

$$\text{Prob}(X = i) = p_{i+} = p_{i1} + p_{i2} + \cdots + p_{in}, \tag{1.7}$$

and the column sums give the distribution of the random variable Y:

$$\text{Prob}(Y = j) = p_{+j} = p_{1j} + p_{2j} + \cdots + p_{mj}. \tag{1.8}$$

Since both of these are probability distributions, they satisfy

$$p_{1+} + p_{2+} + \cdots + p_{m+} = p_{+1} + p_{+2} + \cdots + p_{+n} = 1.$$

Equivalently, the vector of row sums $(p_{1+}, p_{2+}, \ldots, p_{m+}) \in [m]$ lies in the standard $(m-1)-$ simplex Δ_{m-1}, and the vector of column sums $(p_{+1}, p_{+2}, \ldots, p_{+n}) \in [n]$ lies in the standard $(n-1)-$simplex Δ_{n-1}. This means that the linear map

$$\begin{array}{rccc} \alpha : & \mathbb{R}^{m \times n} & \to & \mathbb{R}^m \times \mathbb{R}^n \\ & P = (p_{ij}) & \mapsto & ((p_{1+}, \ldots, p_{m+}), (p_{+1}, \ldots, p_{+n})) \end{array} \tag{1.9}$$

restricts to a projection of convex polytopes

$$\alpha : \Delta_{mn-1} \to \Delta_{m-1} \times \Delta_{n-1}.$$

Definition 1.3.1. *Two random variables X and Y are* **independent**, *written $X \perp\!\!\!\perp Y$, if*

$$\text{Prob}(X = i, Y = j) = \text{Prob}(X = i) \cdot \text{Prob}(Y = j) \text{ for all } (i, j) \in [m] \times [n]. \tag{1.10}$$

The model $\mathcal{M}_{X \perp\!\!\!\perp Y} \subset \Delta_{mn-1}$ of all possible joint probability distributions of two independent random variables is called **independence model**.

Remark 1.3.2. Given the joint probability distribution $P = (p_{ij})$ of two random variables X and Y and taking in mind formulas (1.7) and (1.8), the condition of independence (1.10) of X and Y can be equivalently stated as

$$p_{ij} = p_{i+} \cdot p_{+j} \text{ for all } (i, j) \in [m] \times [n]. \tag{1.11}$$

Thus, the model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ can be defined as

$$\mathcal{M}_{X \perp\!\!\!\perp X} = \{P \in \Delta_{mn-1} : P \text{ satisfies } (1.11)\}.$$

Example 1.3.3. From Remark 1.3.2 we obtain that $\mathcal{M}_{X \perp\!\!\!\perp Y} = Z_\Delta(S)$ with

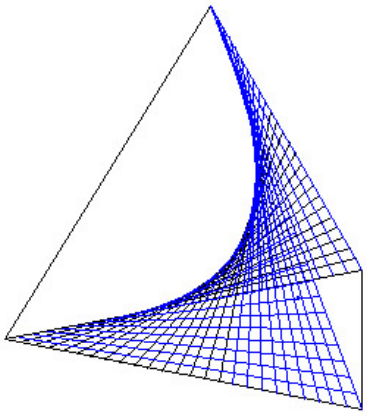
$$S = \left\{ \; p_{ij} - \left(\sum_{k=1}^{n} p_{ik}\right) \cdot \left(\sum_{l=1}^{m} p_{lj}\right) \; \right\}$$

then $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is an algebraic statistical model.

Consider two independent random variables with $m = n = 2$. The model is

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \{(p_{11}, p_{12}, p_{21}, p_{22}) : p_{11}p_{22} = p_{21}p_{12}\} \cap \Delta_3$$

which can be represented as a surface in Δ_3 as you can see in the next figure.



Let us prove now an important result which characterize points in Δ_{mn-1} representing distribution of two independent variables.

Lemma 1.3.4. *Two random variables X and Y are independent if and only if their joint distributions matrix $P = (p_{ij})$ has rank 1.*

Proof. An $m \times n-$matrix $P = (p_{ij})$ has rank 1 if and only if there exist real vectors $(u_1, \dots, u_m)$ and $(v_1, \dots, v_n)$ such that $p_{ij} = u_i v_j$ for all i and j. Here we can scale the entries so that $\sum_{i=1}^{m} u_i = 1$ holds. These conditions imply

$$v_j = \left(\sum_{i=1}^{m} u_i\right) \cdot v_j = \sum_{i=1}^{m} p_{ij} = p_{+j}.$$

Assuming that P lies in Δ_{mn-1}, we get

$$u_i = u_i \cdot \left(\sum_{j=1}^{n} p_{+j}\right) = \sum_{j=1}^{n} u_i v_j = \sum_{j=1}^{n} p_{ij} = p_{i+}.$$

Thus a joint distribution matrix P has rank 1 if and only if $p_{ij} = p_{i+} \cdot p_{+j}$ for all $(i, j) \in [m] \times [n]$, which is precisely, by Remark 1.3.2, the defining condition for X and Y to be independent. □

Remark 1.3.5. Due to the previous Lemma, the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ can be seen as the subset of Δ_{mn-1} consisting of all rank 1 matrices.

In Example 1.3.3 we saw that $\mathcal{M}_{X \perp\!\!\!\perp Y}$ can be defined, as algebraic model, by

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = Z_{\Delta}\Big(p_{ij} - \left(\sum_{k=1}^{n} p_{ik}\right) \cdot \left(\sum_{l=1}^{m} p_{lj}\right)\Big)$$

but these polynomials do not generate $I(\mathcal{M}_{X \perp\!\!\!\perp Y})$. Let $P = (p_{ij})$ be an $m \times n$ matrix of unknowns and let $R = \mathbb{R}[p_{11}, p_{12}, \dots, p_{mn}]$ be the ring polynomials in these unknowns. Let $I_2(P)$ be the ideal generated by all the $2 \times 2-$minors of P. Thus $I_2(P)$ is generated by $\binom{m}{2} \cdot \binom{n}{2}$ quadratic polynomials of the form $p_{ij}p_{kl} - p_{il}p_{kj}$. Then, by Remark 1.3.5 $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is the intersection of the affine variety of $I_2(P)$ with the probability simplex Δ_{mn-1}. Thus $I(\mathcal{M}_{X \perp\!\!\!\perp Y}) =< 2 \times 2 \text{ minors of } P > + < \sum_{i,j} p_{ij} - 1 >$.

This prove the first part of the statement of the following

Proposition 1.3.6. *The independence model $\mathcal{M}_{X \perp\!\!\!\perp Y} \subset \Delta_{mn-1}$ is an algebraic variety of dimension $(m-1)(n-1)$, known as the* **Segre variety***, in the simplex Δ_{mn-1}.*

The computation of the dimension of $\mathcal{M}_{X \perp\!\!\!\perp Y}$ easily follow from the next Lemma.

Lemma 1.3.7. *The restriction of the map α to the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is a bijection between $\mathcal{M}_{X \perp\!\!\!\perp Y}$ and $\Delta_{m-1} \times \Delta_{n-1}$. Its inverse is given by*

$$\Big((u_1, \dots, u_m), (v_1, \dots, v_n)\Big) \mapsto P = (u_i v_j)_{i \in [m], j \in [n]}. \tag{1.12}$$

We denote this bijection $\mathcal{M}_{X \perp\!\!\!\perp Y} \to \Delta_{m-1} \times \Delta_{n-1}$ by the Greek letter μ, i.e. $\mu = \alpha|_{\mathcal{M}_{X \perp\!\!\!\perp Y}}$ and we call it the **moment map of the independence model**. The inverse of the moment map μ^{-1} is given by the formula in (1.12). It maps the convex polytope $\Delta_{m-1} \times \Delta_{n-1}$ bijectively onto the non-negative part of the Segre variety.

Remark 1.3.8. From the map (1.12) we get that $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is a parametric algebraic statistical model where

$$\Theta = \Delta_{m-1} \times \Delta_{n-1}$$

and

$$p_{ij} = f_{ij}(v, u) = v_i u_j.$$

Example 1.3.9. An element of $\mathcal{M}_{X \perp\!\!\!\perp Y}$ can be expressed as a matrix of rank 1, thus

$$Sec^r(\mathcal{M}_{X \perp\!\!\!\perp Y}) = \{\text{matrices } M \in \Delta_{mn-1} \text{ of rank } \leq r\}.$$

For the moment we considered only probabilistic matrices, that is matrices where the sum of the entries is egual to 1. However we can also consider more general matrices and we can eventually transform them in probabilistic matrices simply normalizing them.

Definition 1.3.10. *An $(m \times n)-$matrix with non-negative integer entries is called a* **contingency table**.

Contingency tables are the most basic data structure used by statisticians to record cross-classified discrete data. Every probabilistic matrix is a contingency table. Here is a simple example of such a table. These data concern eye color and hair color of 592 subjects. taken from [49] and discussed in [25].

		Hair Color				
		Black	Brunette	Red	Blonde	**Total**
	brown	68	119	26	7	**220**
	blue	20	84	17	94	**215**
Eye Color	hazel	15	54	14	10	**93**
	Green	5	29	14	16	**64**
	Total	**108**	**286**	**71**	**127**	**592**

A basic statistical question would be whether Hair Color and Eye Color are independent, and, if not, what is the nature of their correlation. These questions can be phrased in the setting of the independence model introduced previously. We regard the eye color as a random variable X whose four values are indexed by the first four positive integers:

$$1 = \text{Brown}, 2 = \text{Blue}, 3 = \text{Hazel}, 4 = \text{Green}.$$

Likewise, hair color is a random variable Y which takes four values:

$$1 = \text{Black}, 2 = \text{Brunette}, 3 = \text{Red}, 4 = \text{Blonde}.$$

If we divide the above $4 \times 4-$matrix by the grand total $N = 592$, which is the sample size of our data, then we get the **empirical distribution**

$$P = \frac{1}{592}\begin{pmatrix} 68 & 119 & 26 & 7 \\ 20 & 84 & 17 & 94 \\ 15 & 54 & 14 & 10 \\ 5 & 29 & 14 & 16 \end{pmatrix} \in \Delta_{15}.$$

Proposition 1.3.11. *The maximum likelihood estimate for an $m \times n-$table is*

$$N \cdot \mu^{-1}(\alpha(P))$$

where P is the empirical distribution for that table and N is the grand total.

Thus the maximum likelihood estimate of a contingency table is the unique rank one table which has the same row and column sums. If A is an $m \times n-$table then we write $\hat{A}$ for its maximum likelihood estimate. For instance, if A is the above table of eye color and hair color data, then

$$\begin{aligned} \hat{A} &= 592 \cdot \mu^{-1}\left(\left(\tfrac{220}{592}, \tfrac{215}{592}, \tfrac{93}{592}, \tfrac{64}{592}\right), \left(\tfrac{108}{592}, \tfrac{286}{592}, \tfrac{71}{592}, \tfrac{127}{592}\right)\right) \\ &= \begin{pmatrix} \frac{1485}{37} & \frac{7865}{74} & \frac{3905}{148} & \frac{6985}{148} \\ \frac{5805}{148} & \frac{30745}{296} & \frac{15265}{592} & \frac{27305}{592} \\ \frac{2511}{148} & \frac{13299}{296} & \frac{6603}{592} & \frac{11811}{592} \\ \frac{432}{37} & \frac{1144}{37} & \frac{284}{37} & \frac{508}{37} \end{pmatrix} = \begin{pmatrix} 40.14 & 106.3 & 26.39 & 47.20 \\ 39.22 & 103.9 & 25.79 & 46.12 \\ 16.97 & 44.93 & 11.15 & 19.95 \\ 11.68 & 30.92 & 7.676 & 13.73 \end{pmatrix} \end{aligned} \tag{1.13}$$

Corollary 1.3.12. *An $m \times n-$contingency table A is independent if and only if it is equal to its own maximum likelihood estimate $\hat{A}$.*

Obviously, we can generalize to tensors all we said in this section. In particular, if we consider n random variables $X_1, X_2, \ldots, X_n$ assuming values respectively in $[m_1], [m_2], \ldots, [m_n]$ then we can define the **probability tensor** $P = (p_{i_1 i_2 \cdots i_n})$ as

$$p_{i_1 i_2 \cdots i_n} = \text{Prob}(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n).$$

Here P is a $(m_1 \times m_2 \times \cdots \times m_n)-$tensor and it represents a point in the probability simplex $\Delta_{m_1 m_2 \cdots m_n - 1}$. Again, the model $\mathcal{M}_{X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_n}$ of n independent random variables can be seen as the subset of $\Delta_{m_1 m_2 \cdots m_n - 1}$ consisting of all rank 1 tensors.

1.4 Linear and toric models

In this section we introduce two classes of models which, under weak conditions on the data, have the property that the likelihood function has exactly one local maximum $\hat{\theta} \in \Theta$. Since the parameter spaces of the models are convex, the maximum likelihood estimate $\hat{\theta}$ can be computed using any of the hill-climbing methods of convex optimization, such as the gradient ascent algorithm.

Definition 1.4.1. *An parametric algebraic statistical model $f : \mathbb{R}^d \to \mathbb{R}^m$ is called a* **linear model** *if each of its coordinate polynomials $f_i(\theta)$ is a linear function, i.e.*

$$f_i(\theta) = \sum_{j=1}^{d} a_{ij}\theta_j + b_i. \tag{1.14}$$

The following proposition states the uniqueness of local maximum $\hat{\theta}$ for this kind of models.

Proposition 1.4.2. *For any linear model Φ and data $\mathbf{u} \in \mathbb{N}^m$, the log-likelihood function $\ell(\theta)$ is concave. If the linear map is one-to-one and all u_i are positive then the log-likelihood function is strictly concave.*

Proof. Our assertion that the log-likelihood function $\ell(\theta)$ is concave states that the Hessian matrix $\left(\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k}\right)$ is negative semi-definite for every $\theta \in \Theta$. In other words, we need to show that every eigenvalue of this symmetric matrix is non-positive. The partial derivative of the linear function $f_i(\theta)$ in (1.14) with respect to the unknown θ_j is the constant a_{ij}. Hence the partial derivative of the log-likelihood function $\ell(\theta)$ equals

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{m} \frac{u_i a_{ij}}{f_i(\theta)}. \tag{1.15}$$

Taking the derivative again, we get the following formula for the Hessian matrix

$$\left(\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k}\right) = -A^T \cdot \operatorname{diag}\left(\frac{u_1}{f_1(\theta)^2}, \frac{u_2}{f_2(\theta)^2}, \ldots, \frac{u_m}{f_m(\theta)^2}\right) \cdot A. \tag{1.16}$$

Here A is the $m \times d$ matrix whose entry in row i and column j equals a_{ij}. This shows that the Hessian (1.16) is a symmetric $d \times d$ matrix each of whose eigenvalues is non-positive.

The argument above shows that $\ell(\theta)$ is a concave function. Moreover, if the linear map f is one-to-one then the matrix A has rank d. In that case, provided all u_i are strictly positive, all eigenvalues of the Hessian are strictly negative, and we conclude that $\ell(\theta)$ is strictly concave for all $\theta \in \Theta$. □

Remark 1.4.3. The critical points of the likelihood function $\ell(\theta)$ of the linear model f are the solutions to the system of d equations in d unknowns which are obtained by equating (1.15) to zero. What we get are the likelihood equations

$$\sum_{i=1}^{m} \frac{u_i a_{i1}}{f_i(\theta)} = \sum_{i=1}^{m} \frac{u_i a_{i2}}{f_i(\theta)} = \cdots = \sum_{i=1}^{m} \frac{u_i a_{id}}{f_i(\theta)} = 0.$$

Our second class of models with well-behaved likelihood functions are the **toric models**. These are also known as **log-linear models**, and they form an important class of exponential families. Let $A = (a_{ij})$ be a non-negative integer $d \times m$ matrix with property that all columns sums are equal:

$$\sum_{i=1}^{d} a_{i1} = \sum_{i=1}^{d} a_{i2} = \cdots = \sum_{i=1}^{d} a_{im} \tag{1.17}$$

The $j-$th column vector a_j of the matrix A represents the monomial

$$\theta^{a_j} = \Pi_{i=1}^{d} \theta_i^{a_{ij}} \quad \text{for } j = 1, 2, \ldots, m$$

Our assumption (1.17) says that these monomials all have the same degree.

Definition 1.4.4. *The* **toric model** $\mathcal{M}_A$ *of* A *is the image of the orthant* $\Theta = \mathbb{R}^d_{>0}$ *under the map*

$$f_A : \mathbb{R}^d \to \mathbb{R}^m, \quad \theta \mapsto \frac{1}{\sum_{j=1}^{m} \theta^{a_j}} \cdot (\theta^{a_1}, \theta^{a_2}, \ldots, \theta^{a_m}). \tag{1.18}$$

Note that we can scale the parameter vector without changing the image: $f_A(\theta) = f_A(\lambda \cdot \theta)$. Hence the dimension of the toric model $f_A(\mathbb{R}^d_{>0})$ is at most $d - 1$. In fact, the dimension of $f_A(\mathbb{R}^d_{>0})$ is one less than the rank of A. The denominator polynomial $\sum_{i=1}^{m} \theta^{a_j}$ is known as the **partition function**. Sometimes we are also given positive constants $c_1, \ldots, c_m > 0$ and the map, now denoted $f_{A,c}$ is modified as follows:

$$f_{A_c} : \mathbb{R}^d \to \mathbb{R}^m, \quad \theta \mapsto \frac{1}{\sum_{j=1}^{m} c_j \theta^{a_j}} \cdot (c_1\theta^{a_1}, c_2\theta^{a_2}, \ldots, c_m\theta^{a_m}). \tag{1.19}$$

In this lecture we consider only the case of $c_i = 1$ for all $i = 1, \ldots, d$.

In a toric model, the logarithms of the probabilities are linear functions in the logarithms of the parameters θ_i. It is for this reason that statisticians refer to toric models as log-linear models.

Maximum likelihood estimation for the toric model (1.18) means solving the following optimization problem

$$\text{Maximize } p_1^{u_1} \cdots p_m^{u_m} \text{ subject to } (p_1 \ldots, p_m) \in f_A(\mathbb{R}^d_{>0}). \tag{1.20}$$

This problem is equivalent to

$$\text{Maximize } \theta^{Au} \text{ subject to } \theta \in \mathbb{R}^d_{>0} \text{ and } \sum_{j=1}^{m} \theta^{a_j} = 1 \tag{1.21}$$

Here we are using multi-index notation for monomials in $\theta = (\theta_1, \ldots, \theta_d)$:

$$\theta^{Au} = \Pi_{i=1}^{d} \Pi_{j=1}^{m} \theta_i^{a_{ij}u_j} = \Pi_{i=1}^{d} \theta_i^{a_{i1}u_1 + a_{i2}u_2 + \cdots + a_{im}u_m} \text{ and } \theta^{a_j} = \Pi_{i=1}^{d} \theta_i^{a_{ij}}.$$

Writing $b = Au$ for the sufficient statistic, our optimization problem (1.21) is

$$\text{Maximize } \theta^{b} \text{ subject to } \theta \in \mathbb{R}^d_{>0} \text{ and } \sum_{j=1}^{m} \theta^{a_j} = 1 \tag{1.22}$$

Proposition 1.4.5. *Fix a toric model A and data $u \in \mathbb{N}^m$ with sample size $N = u_1 + u_2 + \cdots + u_m$ and sufficient statistic $b = Au$. Let $\hat{p} = f(\hat{\theta})$ be any local maximum for the equivalent optimization problems (1.20), (1.21) and (1.22). Then*

$$A \cdot \hat{p} = \frac{1}{N} \cdot b$$

Proof. We introduce a Lagrange multiplier λ. Every local optimum of (1.22) is a critical point of the following function in $d+1$ unknowns $\theta_1, \ldots, \theta_d, \lambda$:

$$\theta^b + \lambda \left(1 - \sum_{j=1}^{m} \theta^{a_j}\right).$$

We apply the scaled gradient operator

$$\theta \cdot \nabla_\theta \left(\theta_1 \frac{\partial}{\partial \theta_1}, \theta_2 \frac{\partial}{\partial \theta_2}, \ldots, \theta_d \frac{\partial}{\partial \theta_d}\right)$$

to the function above. The resulting critical equations for $\hat{\theta}$ and $\hat{p}$ state that

$$(\hat{\theta})^b \cdot b = \lambda \cdot \sum_{j=1}^{m} (\hat{\theta})^{a_j} \cdot a_j = \lambda \cdot A \cdot \hat{p}.$$

This says that the vector $A \cdot \hat{p}$ is a scalar multiple of the vector $b = Au$. Since the matrix A has the vector $(1, 1, \ldots, 1)$)in its row space, and since $\sum_{j=1}^{m} \hat{p}_j = 1$, it follows that the scalar factor which relates the sufficient statistic $b = A \cdot u$ to $A \cdot \hat{p}$ must be the sample size $\sum_{j=1}^{m} u_j = N$. □

Example 1.4.6. Consider the matrix

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

and the vector

$$c = \left(1, \binom{m}{1}, \binom{m}{2}, \ldots, \binom{m}{m-1}, 1\right).$$

Thus the toric model f with components

$$f_j(\theta_1, \theta_2) = c_j \theta^{a_j} = \binom{m}{j} \theta_1^j \theta_2^{m-j}$$

is the model $\mathcal{M}_m$ of Example 1.1.3. Setting $\theta = \frac{\theta_1}{\theta_1 + \theta_2}$ yields "original" parameterization.

1.5 Model invariants

Consider a parametric algebraic statistical model $\mathcal{M}_f = f(\Theta) \subset \mathbb{R}^m$, given by a (parametric) map

$$\begin{array}{rccc} f: & \Theta \subseteq \mathbb{R}^d & \to & \mathbb{R}^m \\ & \theta = (\theta_1, \ldots, \theta_d) & \mapsto & (f_1(\theta), \ldots, f_m(\theta)) \end{array}$$

We first extend the map f to $\mathbb{C}$,

$$\tilde{f} : \mathbb{C}^d \to \mathbb{C}^m$$

then we define the variety $V_{\mathcal{M}_f} := \overline{\tilde{f}(\mathbb{C}^d)}$, that is the zariski closure of the image of $\tilde{f}$.

Let $I_f \subset \mathbb{C}[p_1, \ldots, p_m]$ be the ideal of $V_{\mathcal{M}_f}$: for the Hilbert basis theorem, I_f is finitely generated

$$I_f =< f_1, \ldots, f_r > .$$

Definition 1.5.1. *The ideal I_f is called the* **invariant ideal** *of the model $\mathcal{M}_f$. Its generators $f_1, \ldots, f_r$ are the* **model invariants** *of $\mathcal{M}_f$.*

Example 1.5.2. Since the p_i's are probability, we know that $\sum_{i=1}^m p_i = 1$, i.e. the probability vector $(p_1, \ldots, p_m)$ is a zero of

$$\sum_{i=1}^{m} p_i - 1 = 0$$

This invariant is called **stochastic invariant**.

The variety $V_{\mathcal{M}_f}$ is called the **variety associated to the model** $\mathcal{M}_f$. Varieties are good approximations to parameterizations as stated by the following

Theorem 1.5.3. *$V_{\mathbb{C}}(I(\tilde{f}(\mathbb{C}^d)) \setminus \tilde{f}(\mathbb{C}^d)$ has dimension strictly less than* $\dim(\tilde{f}(\mathbb{C}^d))$.

Remark 1.5.4. It is easy to check that the previous Theorem fails over $\mathbb{R}$. Consider for example the function

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^2$$

Then $f(\mathbb{R}) = [0, \infty]$ but $V_{\mathbb{R}}(I(f(\mathbb{R}))) = \mathbb{R}$.

Example 1.5.5. The variety associated to the toric model $\mathcal{M}_{A,c}$ is defined as $V_{A,c} = \overline{\tilde{f}_{A,c}(\mathbb{C}^d)}$ The invariant ideal is given by

$$I_{A,c} =< \underline{p}^{\mathbf{u}} - \underline{p}^{\mathbf{v}} : A\mathbf{u} = A\mathbf{v}, \mathbf{u}, \mathbf{v} \in \mathbb{N}^m > .$$

Understanding the algebraic structure of $\tilde{f}(\mathbb{C}^d)$, $\overline{\tilde{f}(\mathbb{C}^d)}$ and $I(\tilde{f}(\mathbb{C}^d))$ (or better, of $f(\Theta)$, $\overline{f(\Theta)}$ and $I(f(\Theta))$) can be useful for making statistical inference.

Roughly speaking, V_f is a variety that contains the (complex) joint distribution for all possible choices of (complex) numerical parameters $\theta_1, \ldots, \theta_d$ for the model $\mathcal{M}_f$. In applications, the choose of the model representing a given event is usually the element of greatest interest. If an observed probability distribution were "close" to V_f, then $\mathcal{M}_f$ could be interpreted as a model fitting the observed data.

Remark 1.5.6. Extending the parameterization f to the complex numbers from the stochastic setting is done because an algebraically closed field provides the easiest and most natural setting for understanding polynomial maps. Of course, complex parameters and complex joint distributions are not so natural from a statistical viewpoint. Obviously, the final goal would be to understand the model in the stochastic setting.

We have the following method which is statistically consistent.

Algebraic Algorithm for Statistical Inference:

INPUT: the joint distribution of observed frequencies $\hat{P}$.

i) fix a set of models $\Omega = \{\mathcal{M}_1, \mathcal{M}_2, \dots\}$;

ii) for each model $\mathcal{M}_i$ in Ω

- Find some/most/all invariants F for the variety V_i associated to $\mathcal{M}_i$;
- Test if $F(\hat{P}) \approx 0$.

OUTPUT: the model $\mathcal{M}_i$ for which $\hat{P}$ is as close as possible to V_i.

One part of understanding V_f is describing it implicitly, as the zero set of polynomials. This means to find polynomials $F \in \mathbb{C}[p_1, \dots, p_m]$ such that $F(q) = 0$ for all $q \in V_f$. This is equivalent to find the kernel of the map

$$\tilde{\Phi} : \mathbb{C}[p_1, \dots, p_m] \to \mathbb{C}[\theta_1, \dots, \theta_d]$$

The kernel of $\tilde{\Phi}$ is the ideal I of the polynomials in the p_i's vanishing for all choices of (stochastic or complex) parameters θ_i's, i.e. it corresponds exactly to the ideal defined in Definition 1.5.1.

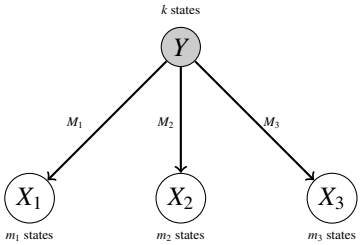
Remark 1.5.7. In the previous sections we considered elements in the unitary simplex Δ_m. This probabilistic condition can be interpreted in Algebraic Geometry as the fact that each probability distribution is an element of a certain affine subspace of a projective space $\mathbb{P}^{m-1}$. At the same time, we can view $V_{\mathcal{M}}$ projectively. In fact, by the stochastic invariant, one has $V_{\mathcal{M}} \subset \mathbb{P}^{m-1}$. The passage to the projective case forces us to look only for phylogenetic invariants among homogeneous polynomials.

What we want to do is to find explicitly the ideal for each parametric algebraic statistical model $\mathcal{M}$. Since, by Hilbert Basis Theorem ([26], Theorem 1.2, page 27), these ideals are finitely generated, this is equivalent to give a list of generators of each ideal. The research of invariants seems to be a specific issue of computational Algebraic Geometry. Here, the main techniques to manipulate polynomials are given

by Gröbner Basis ([26], Chapter 15). Theorethicaly, Gröbner basis permit to find all the invariants of $\mathcal{M}$. Unluckily, in the practical situation the use of Gröbner Basis is limited to a small number of parameters and states. In fact, the basic alghorithm to give I_T is the application of the elimination process to the set of equations $p_i(\theta) - \sum_{a\in\mathbb{N}^d} c_a \theta_1^{a_1}\theta_2^{a_2}\cdots\theta_d^{a^d} = 0$ with respect to the indeterminates/parameters $\theta_1, \ldots, \theta_d$. Thus, as soon as the number of states, the number of parameters or the degree of polynomials grow, the computation becomes more and more complex and technology, at the moment, is not able to produce any results. Hence, one of the main research problem in Algebraic Statistics is to exhibit combinatorial methods to find invariants without using elimination theory. We will see two different cases in Lectures 2 and 4.

1.6 Graphical models

Consider the following graph



where the grey node has a hidden random variable Y with state space $[k]$ and the three white nodes have three observable random variables X_1, X_2 and X_3 with state spaces respectively $[m_1]$, $[m_2]$ and $[m_3]$. At each edge we associate a $k \times a_t-$matrix $M^{(t)}$, $t = 1, 2, 3$, where the $(i, j)-$entry $m_{ij}^{(t)}$ is the probability to pass from state i in Y to state j in X_t. Then each row of M_t has sum 1.

Suppose that the state at Y is momentarily fixed as $\tilde{k}$. Then, for each edge leading away from Y, towards a white node, we have a point $\overline{m}_{\tilde{k}X_t} = (m_{\tilde{k}1}^{(t)}, \ldots, m_{\tilde{k}a_t}^{(t)}) \in \mathbb{P}^{k-1}$ that represents the $\tilde{k}-$th row of the transition matrix $M^{(t)}$. Thus, if we define

$$P^{\tilde{k}} := \overline{m}_{\tilde{k}X_1} \otimes \overline{m}_{\tilde{k}X_2} \otimes \overline{m}_{\tilde{k}X_3} \in \mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \mathbb{P}^{a_3-1}$$

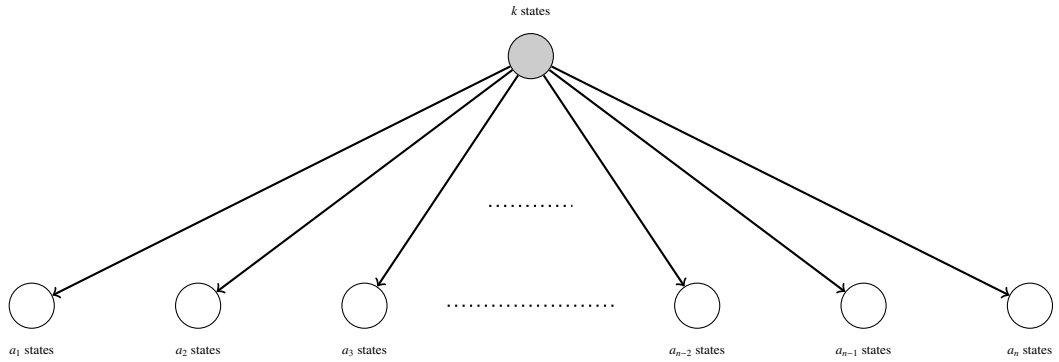
then $P^{\tilde{k}}$ is a point in the Segre product ([35], Example 2.11, page 25) of three projective spaces whose entries (up to scaling) are the expected frequencies of observing pattern at the white nodes conditioned by the state at Y being $\tilde{k}$. Summing over all possible states at Y, we obtain the joint distribution

$$P = P^1 + P^2 + \cdots + P^k.$$

Since we are summing k points on the variety $\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \mathbb{P}^{a_3-1}$, we obtain $P \in V_T = S_k(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \mathbb{P}^{a_3-1})$, i.e. the $k-$secant variety ([35], Example 8.5, page 90) of the Segre product $\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \mathbb{P}^{a_3-1}$.

We have to point out that the distribution of initial states at Y does not explicitly appear, since it has been accounted for in the arbitrary scaling factors that appear in each P^i, when we choose particular projective coordinates to express them. Hence, the joint distribution P has been decomposed as the sum of k rank 1 tensors, one for each possible state at Y (and this forces P to have rank k, by the definition itself of tensor rank).

More generally, we can consider any number n of observable nodes:



The variety associated to this model is the $k-$secant variety of the Segre product $\mathbb{P}^{\alpha_1-1} \times \mathbb{P}^{\alpha_2-1} \times \cdots \times \mathbb{P}^{\alpha_n-1}$.

These kinds of models are specific examples of the class of so called **graphical models**. When $a_1 = a_2 = \cdots = a_n = 2$ the previous graphical model is called **Bernoulli model**.

Lecture 2. Phylogenetic Algebraic Geometry

2.1 Introduction

One of the main problems in modern Biology is that of phylogenetic inference. Let us consider a model of molecular evolution (for example, DNA sequences) and suppose that evolution occurs along a bifurcating tree, proceeding from a root, i.e. the common ancestral species, toward the leaves, i.e. the descendant species. We require that, at each site in the sequences, bases mutate according to a probabilistic process that depends upon the edges of the tree. Only the sequences at the leaves of the tree can be observed, while sequences at internal nodes correspond to hidden variables in this graphical model. Thus, the phylogenetic inference concerns the problem to infer the tree topology from observed sequences, assuming some probabilistic (and reasonable) model.

In 1987, Cavender and Felsenstein [15] and, separately, Lake [43], introduced an algebraic approach to attack this problem. In fact, under many standard models of molecular evolution, for a fixed tree topology, the joint distribution of bases at the leaves are described by polynomial equations in the parameters of the model. They proposed to search for polynomials, called phylogenetic invariants, which vanish on any joint distribution arising from the tree and model, regardless of parameter values, in a similar way, later, the concept of model invariant was introduced in Algebraic Statistics.

Recently, several authors have started to research and study phylogenetic invariants by a deeper use of Algebraic Geometry, also in connection with Algebraic Statistics. Although the idea of Phylogenetic Algebraic Geometry had already been undertaken in their works (for example, [4], [5], [6], [45] and [52]) only in [28] we can find its definition for the first time. Here the authors say that "*Phylogenetic Algebraic Geometry is concerned with certain complex projective algebraic varieties derived from finite trees. By Phylogenetic Algebraic Geometry we mean the study of algebraic varieties which represent statistical*

models of evolution".

The varieties which arise from such a kind of model can be different. We can find, for example, the more familiar ones, as secant varieties, determinantal varieties, toric varieties and Segre–Veronese varieties. This happens, in general, when we consider models related to small trees, i.e. trees with at most five leaves. For trees with more than six leaves instead, we can encounter families of new kinds of varieties, often completely unknown. The study of such varieties is related especially to the search for the generators of their ideals. By the Hilbert Basis Theorem we know that these generators are in a finite number and are exactly the phylogenetic invariants associated to the corresponding tree.

For general background reading on Phylogenetics, we strongly suggest the books by Felsenstein [30] and Semple-Steel [48]. They deeply analyze evolutionary trees according to Biology, Computer Science, Statistics and Mathematics. Instead, for a survey of Algebraic Statistics and Computational Biology, the book, [46], edited by Lior Pachter and Bernd Sturmfels, is surely the best choice. There is large literature about Algebraic Geometry and Commutative Algebra. The elements we need, though, can be found in [26] (Chapter 0), [35] (Lectures 1,2 and 8) and [36] (Chapter 1). For the interested reader, we suggest also books [21] and [22], where the authors introduce concepts and results in Algebra and Geometry with the perspective of possible applications. For references to others research papers we recommend again [30] and [48]. For the most recent ones, the reader can consult [9] and [28]. In [28], there is also a very interesting section where the authors collect a series of open problems.

2.2 Evolutionary trees and Markov models

Since graphs play an important role in Phylogenetics, we will start recalling some basic facts about them.

Definition 2.2.1. *A* **graph** G *is an ordered pair* (V, E) *consisting of a non-empty set* V *of* **vertices** *and a multiset* E *of* **edges** *each of which is an element of* $\{(x, y) : x, y \in V\}$. *An edge that joins a vertex to itself is a* **loop** *and the edges that join the same distinct pair of vertices are called* **parallel edges**.

All the graphs we consider will have a finite set of vertices.

If $e = \{u, v\}$ is an edge of a graph G, then u and v are **adjacent** and e is said to be **incident** with u and v. The vertices u and v are the **ends** of e. Let v be a vertex of a graph G. The **valency** of v, $\nu(v)$, is the number of edges in G that are incident with v. A **path** in a graph G is a sequence of distinct vertices $v_1, v_2, \ldots, v_k$ such that, for all $i = 1, \ldots, k-1$, v_i and v_{i+1} are adjacent. If, in addition, v_1 and v_k are adjacent, then the subgraph of G, whose vertex set is $\{v_1, v_2, \ldots, v_k\}$ and whose edge set is $\{(v_k, v_1)\} \cup \{(v_i, v_{i+1}) : i = 1, \ldots, k-1\}$, is a **cycle**. A graph is **connected** if each pair of vertices in G can be joined by a path: otherwise G is **disconnected**.

Let us denote by $|F|$ the cardinality of a set F. We recall the following

Lemma 2.2.2. *Let $G = (V, E)$ be a graph, then*

$$\sum_{v \in V} \nu(v) = 2|E|.$$

Moreover, if G is connected, one has $|V| \leq |E| + 1$.

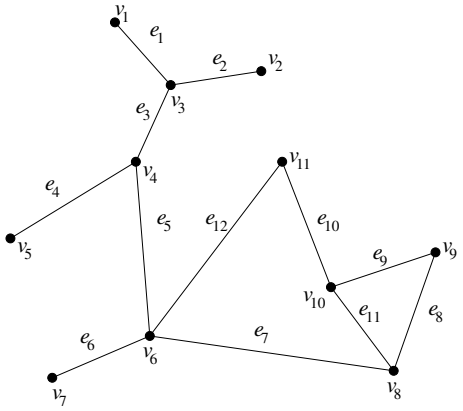


Fig. 1a): A connected graph

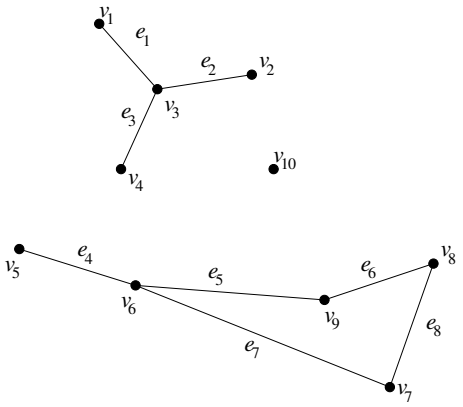


Fig. 1b): A disconnected graph

Graphs have several applications in Biology: *food web* and *competition graphs*, *genome mapping* and *interval graphs*, *pedigree (di)graphs*. Here we deal with another application: the theory of phylogenetic trees.

Definition 2.2.3. *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.*

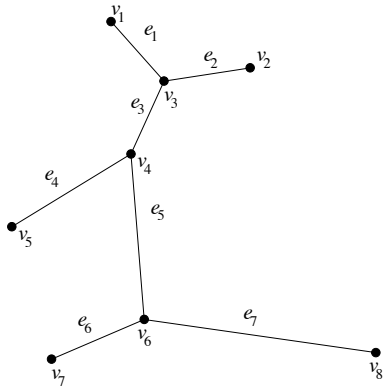


Fig 2: An example of tree

An important characterization of trees is given by

Theorem 2.2.4. *Let $G = (V, E)$ be a graph. Then the following are equivalent:*

1) G is a tree;

2) for any two vertices v and u in V there exists a unique path in G from v to u;

3) G is connected and $|V| = |E| + 1$.

A vertex of a tree of valency one is called a **leaf**. We denote by L the set of leaves and define $\tilde{V} := V \setminus L$ the set of interior vertices. Similarly, we denote by $\tilde{E}$ the set of interior edges. A tree is **binary**, or **bifurcated**, if every interior vertex has valency three. Two distinct leaves of a tree are said to form a **cherry** if they are adjacent to a common vertex. For example, in Figure 2, the pairs $\{v_1, v_2\}$ and $\{v_7, v_8\}$ are cherries.

A **rooted** tree is a tree 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

$$v_i \leq_T v_j \text{ if the path from the root of } T \text{ to } v_j \text{ includes } v_i.$$

In this case we say that v_j is a **descendant** of v_i and that v_i is an **ancestor** of v_j. For this reason we always draw a rooted tree with the root r at the top of the figure and oriented so as to respect the ancestor-descendant relationship.

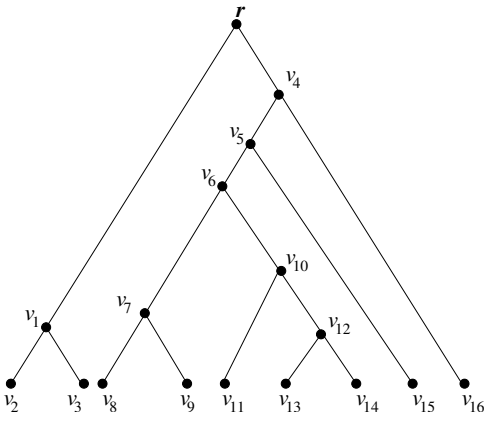


Fig. 3: A rooted tree

Let us state, now, the definition of phylogenetic tree. Among several definitions, we will choose the one closer to Biology:

Definition 2.2.5. *An* **X-tree** *$\mathcal{T}$ is an ordered pair (T, ϕ), where T is a tree with vertex set V, label set X and $\phi : X \to V$ is a map with the property that, for each $v \in V$ of valency at most two, $v \in \phi(X)$. 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 valency three, $\mathcal{T}$ is a **binary phylogenetic tree**.

It is common in Biology to focus on binary trees (i.e., trivalent, except bivalent at the root) as being of primary interest. In fact, most speciation events are believed to be of the sort where only two species at a time arise from a parent species. Multifurcations in a tree might be used to represent ignorance such as when several speciation events occur so closely in time that we are unable to resolve their order.

From now on, if not otherwise specified, we will consider only binary trees.

Proposition 2.2.6.

i) Let $\mathcal{T}$ be a binary phylogenetic X−tree and let $n = |X|$, Then, for all $n \geq 2$, $\mathcal{T}$ has $2n - 3$ edges and $n - 3$ interior edges.

ii) Let $B(n)$ be the set of all binary phylogenetic trees with label set $X = \{1, 2, \ldots, n\}$. If $n = 2$ then $|B(n)| = 1$. If $n \geq 3$ then

$$|B(n)| = \frac{(2n-4)!}{(n-2)!2^{n-2}} = 1 \times 3 \times 5 \times \cdots \times (2n-5)$$

Proof. See [48], Propositions 2.1.3 and 2.1.4. □

Obviously, we can extend the notion of an X−tree to the rooted case.

Definition 2.2.7. *A* **rooted X-tree** $\mathcal{T}$ *is an ordered pair (T, ϕ), where T is a rooted tree with vertex set V, rooted vertex r, label set X and $\phi : X \to V$ is a map with the property that, for each $v \in V \setminus \{r\}$ of valency at most two, $v \in \phi(X)$. A rooted X−tree is also called a* **rooted semi-labelled tree** *(on X). A* **rooted phylogenetic tree** *is a rooted X−tree (T, ϕ) with the property that ϕ is a bijection from X into the set of leaves of T and the root has valency at least two. If, in addition, every interior vertex of T has valency three, $\mathcal{T}$ is a* **rooted binary phylogenetic tree**.

Let $\mathcal{T}$ be a rooted X−tree and let x, y be two leaves. We denote $lca(x, y)$ the most recent common ancestor of x and y. For example, in Figure 3, one has $lca(v_8, v_9) = v_7$, $lca(v_8, v_{11}) = lca(v_8, v_{12}) = v_6$, $lca(v_8, v_{16}) = v_4$. For a rooted phylogenetic tree $\mathcal{T}$ on X, we view the edges of $\mathcal{T}$ as being directed from the root r. Then we consider $\mathcal{T}$ as describing the evolution of the set X of extant species that label the leaves of $\mathcal{T}$ from a common hypothetical ancestral species at r; the other interior vertices of $\mathcal{T}$ correspond to further hypothetical ancestral species or to past speciation events. Thus $lca(x, y)$ can be seen as the most recent shared ancestral species (or speciation event) of the species x and y.

Remark 2.2.8. Unrooted phylogenetic trees are also biologically relevant because they are typically what the tree reconstruction methods generate. We can observe that it is always possible to pass from an unrooted tree to a rooted one and viceversa. In particular, passing from the unrooted to the rooted tree, means to choose an internal vertex as the root or add another vertex inside an edge and choose it as the root.

Remark 2.2.9. In general, as X, we will use the set $\{1, 2, \ldots, n\}$, where each number will correspond to a specific species.

Remark 2.2.10. Let us mention some particular kinds of trees. The (**rooted**) **caterpillar tree** is any (rooted) binary phylogenetic tree for which the induced subtree on the interior vertices is a path graph. A **rooted balanced tree** of height $h \geq 0$ is a rooted binary phylogenetic tree, with $n = 2^h$ leaves, each of which is separated from the root by exactly h edges. A **star tree** is a phylogenetic tree with no interior edges, i.e. with a single interior vertex that is adjacent to all the leaves. In Figures 4 and 5 we show respectively the unrooted and rooted cases. Each rooted case is obtained by adding another vertex inside an edge in the respective unrooted case.

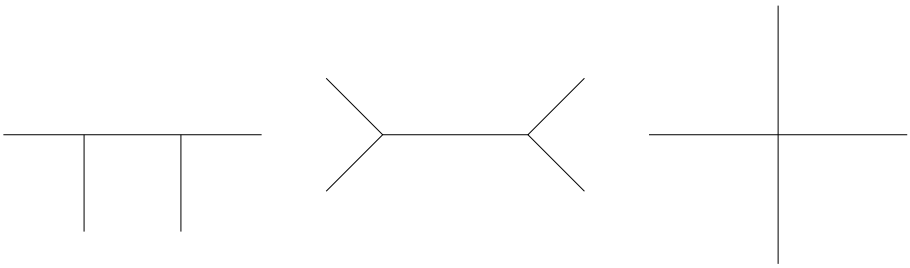


Fig 4 a) - b) - c): A caterpillar tree, a balanced tree of height 2, a star tree

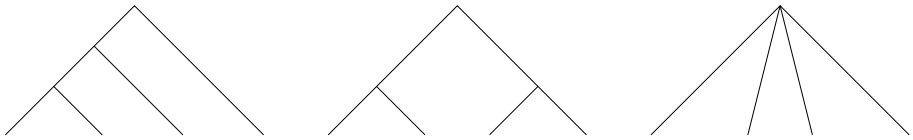


Fig 5 a) - b) - c): The rooted cases

We introduce now the concept of Markov process.

Let $X_1, \ldots, X_t$ be random variables on a sample space S taking value in a set U and let $A = \{1, 2, \ldots, t\}$. For a subset $B \subset A$ and an event E of S we will write $\text{Prob}(E| \cap_{i\in B} \{X_i\})$ for $\text{Prob}(E| \cap_{i\in B} \{X_i = u_i\})$, i.e. the conditional probability of E given $\cap_{i\in B}\{X_i = u_i\}$, for every selection of $u_i \in U$.

We fix an alphabet with k letters, for instance $[k] = \{1, 2, ..., k\}$.

Definition 2.2.11. *Let T be a rooted tree with vertex set V. A Markov process on T, with state set $[k]$, is a*

family $\{X_v : v \in V\}$ *of random variables such that, whenever* (u, v) *is arc of* T*, with* $v < u$*, and* $\alpha \in [k]$*,*

$$\text{Prob}(X_v = \alpha | \cap_{w<v} X_w) = \text{Prob}(X_v = \alpha | X_u) \tag{2.23}$$

Condition (2.23) is known as the **Markov Property**. Intuitevely, this states that, for each arc (u, v) of T, the value of X_v, conditional on X_u, is independent of the $X-$values at all other "earlier" vertices.

Let T be a rooted tree. For each edge $e = (u, v)$ of T (with $u < v$), a Markov process on T, with state set $[k]$, induces an associated $k \times k$ transition matrix, denoted $M^{(e)}$, where the $(i, j)-$entry, $m_{ij}^{(e)}$, is the probability to pass from state i on u to state j on v. We ask for

i) $m_{ij}^{(e)} \geq 0$;

ii) $\sum_{j=1}^{k} m_{ij}^{(e)} = 1$.

Thus, if we specify a Markov matrix for each edge of the tree, we have modeled how the entire evolutionary process proceeds along the tree.

Once we fixed the root r we define also a root distribution

$$\pi(r) = (\pi(r)_1, \pi(r)_2, \ldots, \pi(r)_k)$$

where $\pi(r)_i$ is the probability to have the state i at the root. Obviously $\pi(r)_i \geq 0$, $\forall i = 1, \ldots, k$ and $\sum_{i=1}^{k} \pi(r)_i = 1$. The root distribution vector $\pi(r)$ gives probabilities of the various states for the variable at the root, while $k \times k$ Markov matrices give transition probabilities of state changes from ancestral to descendant nodes along each edge.

Definition 2.2.12. *We refer to data* $(T, \mathcal{M})$ *as the* **general Markov model** *on* T*, where* $\mathcal{M} = (\pi(r), \{M^{(e)} : e \in E\})$*. We often refer to* $\mathcal{M}$ *as* **stochastic parameters***, distinguishing them from tree parameters.*

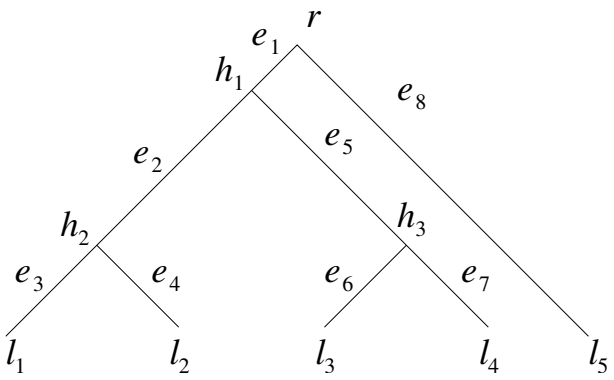
Remark 2.2.13. As to the DNA, the number of states is $k = 4$, but as to protein sequences, which are built from twenty amino acids, $k = 20$. The case $k = 2$ is also of interest for DNA substitution models, if we group bases into purines R = {A, G} and pyrimidines Y = {C, T}.

Let $l_1, \ldots l_n$ be the leaves of the tree T. Evolution occurs along the tree, but we can observe sequences only at the leaves of T. With the parameters of the model $\mathcal{M}$ thus specified, we are interested in the **joint distribution** P **of states at the leaves** l_i's. The joint distribution P is an $n-$dimensional $k \times k \times \cdots \times k$ tensor (or table or array) with entries

$$P(i_1, \ldots, i_n) = \text{Prob}(l_1 = i_1, \cdots, l_n = i_n)$$

where Prob($l_1 = i_1, \cdots, l_n = i_n$) represents the probability to have state i_j at the leaf l_j, for $j = 1, \ldots, n$. In general, we will denote $P(i_1, \ldots, i_n)$ by $p_{i_1 \cdots i_n}$. The entries of P are the expected frequencies to be seen in patterns of states $(i_1, \ldots, i_n)$ at the leaves of the tree. These expected pattern frequencies can be explicitly expressed in terms of the parameters of the model, as we can explain through an example.

Example 2.2.14. Consider the tree with leaves $l_1, \ldots, l_5$.



Let $M^{(e)} = (m^{(e)})_{ij}$ be the $k \times k$ matrix on the edge h_e, for $e = 1, \ldots, 8$, and $\pi(r)$ the root distribution. Suppose that we want to compute the probability $p_{i_1 \cdots i_5}$. If we start from a state w_0 at the root, we can see that $\pi(r)_{w_0} m^{(1)}_{w_0, w_1}$ is the probability to have a state w_1 at the vertex h_1. Moving in this way, we can see that we reach leaf l_1 with state i_1 by

$$\pi(r)_{w_0} m^{(1)}_{w_0,w_1} m^{(2)}_{w_1,w_2} m^{(3)}_{w_2,i_1}$$

where w_2 is an unobserved state at the vertex h_2. The procedure for the leaf l_2 is similar, but, since we already have the probability of transition until vertex h_2, from the computation on l_1, it is enough to multiply the previous term by $m^{(4)}_{w_2,i_2}$. Now it is clear how to proceed. Thus we obtain

$$\pi(r)_{w_0} m^{(1)}_{w_0,w_1} m^{(2)}_{w_1,w_2} m^{(3)}_{w_2,i_1} m^{(4)}_{w_2,i_2} m^{(5)}_{w_1,w_3} m^{(6)}_{w_3,i_3} m^{(7)}_{w_3,i_4} m^{(8)}_{w_0,i_5}$$

This is the probability to have state i_j at the leaf l_j, $j = 1, \ldots, 5$, and states w_0, w_1, w_2 and w_3 respectively at the root r and at vertices h_1, h_2, h_3. Since the internal nodes are hidden, we have to consider all possible states at the internal nodes, thus the final probability will be

$$p_{i_1 \cdots i_5} = \sum_{\substack{1 \le w_i \le k \\ i = 0,1,2,3}} \pi(r)_{w_0} m^{(1)}_{w_0,w_1} m^{(2)}_{w_1,w_2} m^{(3)}_{w_2,i_1} m^{(4)}_{w_2,i_2} m^{(5)}_{w_1,w_3} m^{(6)}_{w_3,i_3} m^{(7)}_{w_3,i_4} m^{(8)}_{w_0,i_5}$$

In general, let $T = (V, E)$ be an $n-$taxon tree with Markov model $\mathcal{M} = (\pi(r), M^{(e)})$. Let us denote by $s(e)$ and $f(e)$ the ends of e. Thus the joint distribution P is given by the formula

$$P(i_1, \ldots, i_n) = \sum_{(b_v)\in H} \left[\pi(r)_{b_r} \prod\nolimits_e \left(m^{(e)}_{b_{s(e)}, b_{f(e)}} \right) \right] \tag{2.24}$$

where the product is taken over all edges e getting away from the root r and the sum is taken over the set

$$H = \{(b_v)_{v\in V} | b_v \in [k] \text{ if } v \neq i_j, b_v = i_j \text{ if } v = i_j\} \subset [k]^{2n-2}.$$

We can say that H represents the set of all "histories" consistent with the specified states at the leaves. More generally, for the general Markov model on an $n-$taxon tree, each probability $p_{i_1,\ldots,i_n}$ will be a degree $2n-2$ polynomial, with k^{n-2} terms. The precise form of these polynomials reflects the topology of the tree T.

Remark 2.2.15. The model we have described here concerns a base substitution process at a single site. In general, for phylogenetic inference, the data are aligned DNA sequences of some length L. Thus, we assume that the evolutionary process at each site proceeds independently of all other sites, but according to the same probabilistic process, with the same parameters. This independent, identically distributed (i.i.d.) assumption is not desirable from a biological viewpoint. In fact, we can have substitutions at one site which are not independent. However, a form of the i.i.d. assumption is essential since only by viewing each site as a trial of the same process, we can obtain enough data to infer something about the parameters.

Let us point out the following important

Proposition 2.2.16. *Fix an $n-$taxon tree T. Let r be a choice of root for T (which may be a leaf, an internal node of valency 3, or along some edge). Then, for a generic choice of stochastic parameters S_r for the general Markov model rooted at r, and for any other choice of a root q for T, on either a leaf or an internal node of valency 3, there is a uniquely determined choice of general Markov model parameters S_q for the model rooted at q producing the same joint distribution at the leaves as S_r.*

Proof. See [4], Proposition 1. □

A consequence of the previous Proposition is that the location of the root in a tree T is a biological problem, not a mathematical one.

The general Markov model has more parameters than models typically use in practice. Once the tree parameter has been chosen as a particular $n-$taxon tree, there are $k-1$ free choices on $\pi(r)$ (since $\sum_{i=1}^{k} \pi(r)_i = 1$) and $k(k-1)$ free choices on the entries of the matrix $M^{(e)}$, for each edge e. Thus, one has $N := (2n-3)k(k-1) + k - 1$ numerical parameters. The growth is only linear in the number of taxa, but the coefficient,

depending on k, could be very large. For example, for $k = 2$, the total number of parameters is $4n - 2$, while, for $k = 4$, it grows as $24n$. The number of parameters has several effects on the inference: slow computations, overfitting. If the data can be described by a model with fewer parameters, that model may provide a better basis for inference. Thus, in general, we consider particular restrictions on the stochastic parameters of the general Markov model, given by mathematical and/or biological reasons.

Let us introduce now some examples of submodels of the general Markov model. The reader can find a wider range of submodels in [7] (the *General Time Reversible* model and *Mixture* model) and in [5] (the *Stable Base Distribution* model, the *Simultaneous Diagonalization* model, the *Algebraic Time Reversible* model).

2.2.1 The Jukes-Cantor model for DNA

This model is the biologically-plausible model with the fewest parameters. It assumes a uniform root distribution vector $\pi(r) = (0.25, 0.25, 0.25, 0.25)$ and edge transition matrices of the form

$$M^{(e)} = \begin{pmatrix} 1 - a_e & \frac{a_e}{3} & \frac{a_e}{3} & \frac{a_e}{3} \\ \frac{a_e}{3} & 1 - a_e & \frac{a_e}{3} & \frac{a_e}{3} \\ \frac{a_e}{3} & \frac{a_e}{3} & 1 - a_e & \frac{a_e}{3} \\ \frac{a_e}{3} & \frac{a_e}{3} & \frac{a_e}{3} & 1 - a_e \end{pmatrix}$$

where a_e could vary for each edge e.

2.2.2 The Kimura 2-parameter model

Because of chemical similarities, the bases are classified as purines {A, G} and pyrimidines {C, T}. Assigning probability a to in-class changes (transitions), and b to out-of-class changes (transversions), we arrive at the Kimura 2-parameter model, with matrices

$$M^{(e)} = \begin{pmatrix} 1 - (a_e + 2b_e) & a_e & b_e & b_e \\ a_e & 1 - (a_e + 2b_e) & b_e & b_e \\ b_e & b_e & 1 - (a_e + 2b_e) & a_e \\ b_e & b_e & a_e & 1 - (a_e + 2b_e) \end{pmatrix}$$

where the rows and columns are ordered by the states A, G, C, T (purines, followed by pyrimidines). Typically $a > b$, since transitions are often observed more frequently than transversions.

2.2.3 The Kimura 3-parameter model

A slight generalization, introduced more for its mathematical structure than for biological reasons, is the Kimura 3-parameter model with transition matrices of the form

$$M^{(e)} = \begin{pmatrix} 1-\rho_e & a_e & b_e & c_e \\ a_e & 1-\rho_e & c_e & b_e \\ b_e & c_e & 1-\rho_e & a_e \\ c_e & b_e & a_e & 1-\rho_e \end{pmatrix}$$

where $\rho_e = a_e + b_e + c_e$. A fundamental result on these structures is given by Hadamard conjugation ([38], [39]) and it permits to introduce Fourier analysis as a tool for studying such models.

Remark 2.2.17. Many probabilistic models of the mutation process - as evolution proceeds down a tree - focus on a single site in a sequence, and only on base substitutions occurring at that site. In general we introduce more complicated models when we want to consider different types of sequence changes as insertions, deletions and inversions ([45]).

2.3 Phylogenetic invariants

In 1987, Cavender and Felsenstein in [15], and, separately, Lake in [43] introduced the concept of phylogenetic invariants as a new approach to the study of phylogenetic trees arising from biological sequence data (i.e. the case of $k = 4$ states: A,C,G,T). Obviously, we just consider the generalization to the case of k states, with k an arbitrary positive integer, $k \geq 2$.

We recall that, given a topological tree T with n leaves (or terminal taxa) and a model $\mathcal{M}$ of evolution along this tree, it is possible to compute the expected pattern frequencies of the k^n patterns of various states at the leaves, in terms of the parameters of the model, as explained in (2.24).

Definition 2.3.1. *A* **phylogenetic invariant***, for the topological tree T and the parameterized model $\mathcal{M}$, is a polynomial in k^n variables, which becomes zero when the expected frequencies are substituted for the variables.*

Since we want to consider an algebraic approach, we can work over the complex field. Thus, we will talk of **complex parameters** to distinguish them from **stochastic parameters**, that is, positive real numbers. In both cases, we require that the root distribution and each row of the transition matrices sum to 1.

Let $\{z_{i_1 \cdots i_n}\}$ be a set of k^n indeterminates indexed by $1 \leq i_1, \ldots, i_n \leq k$, and denote by R the polynomial ring $\mathbb{C}[z_{i_1 \cdots i_n}]$. We can restate the previous definition in the following way.

Definition 2.3.2. *A phylogenetic invariant, for the general Markov model* $(T, \mathcal{M})$*, is a polynomial* $f \in \mathbb{C}[z_{i_1\cdots i_n}]$ *such that* $f \equiv 0$ *under the substitution* $p_{i_1\cdots i_n} \to z_{i_1\cdots i_n}$ *of the polynomial expressions for the expected pattern frequencies at the leaves.*

Example 2.3.3. Since the $p_{i_1\cdots i_n}$'s represent all the possible probabilities of the events in the joint distribution state, one has

$$\sum_{1\le i_1,\dots,i_n\le k} p_{i_1\cdots i_n} = 1 \tag{2.25}$$

This invariant, which is common to all $n-$taxon trees with k states, is called **stochastic invariant**.

Suppose that phylogenetic invariants can be found. This permits us to choose both the topological tree and the pararameterized model. In fact, starting from the observed data, we can compute the observed frequencies of patterns $\hat{P}_{i_1\cdots i_n}$'s. The observed data are distinct sequences of states (in particular, in the case $k = 4$, these are aligned DNA sequences), one for each of the n species. All sequences have the same length M. Thus, the observed frequencies of patterns $\hat{P}_{i_1\cdots i_n}$'s are given by

$$\hat{P}_{i_1\cdots i_n} = \frac{\text{occurrence of } i_1,\dots,i_n}{M}.$$

Example 2.3.4. Consider four species with given DNA sequences of length 30.

Species 1	CG**TT**ACCCACTAGTTTA**T**GACG**TT**ACCCAC
Species 2	CG**TT**ACCGACTAAATGC**T**GTCG**TT**ACCGAC
Species 3	AG**CC**CCCCCAATTATGAG**C**GTAG**CC**CCCCCAA
Species 4	CG**GG**ATTAAAATGCCGC**G**GGCG**GG**ATTAAA

Thus, for example, one has $\hat{P}_{TTCG} = \frac{5}{30} = 0.16667$.

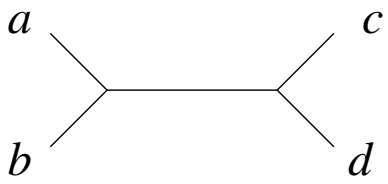
If observed frequencies of patterns are good estimators of the expected frequencies, they will force the invariants to vanish or, at least, to be small. Thus, we choose the topological tree so that its invariants are close to vanish on the observed frequencies (then, in order to apply invariants to real data, one must decide what it means for an invariant to be "close to vanishing" on observed frequencies). More precisely, consider a phylogenetic invariant f for a general Markov model $(T, \mathcal{M})$, where $\mathcal{M}$ is given in the unknown parameters $\pi(r)_l, m_{ij}^{(e)}$. Let $P = (p_{i_1\cdots i_n})$ be the joint distribution tensor describing probabilities of states at the leaves. Hence $f(P) = 0$. These probabilities are expressed in terms of the parameters of the model $\mathcal{M}$, that is, the entries $m_{ij}^{(e)}$ and the root distribution. Replacing P with a joint distribution tensor P_0, arising from any specific choice of parameters for T and $\mathcal{M}$, one has again $f(P_0) = 0$. Call $\hat{P}$ the tensor representing the observed pattern frequencies of the data. If T and $\mathcal{M}$ are the correct tree and model relating to the sequence,

then $\hat{P} \approx P_0$. Since $f(P_0) = 0$, then $f(\hat{P}) \approx 0$. Thus, the near vanishing of the phylogenetic invariants on the observed pattern frequencies is a good verification on T and $\mathcal{M}$ as correct tree and model.

This model-based method of choosing topological trees will be useful if we are able, first of all, to produce "efficiently" phylogenetic invariants. Several techniques are used to find phylogenetic invariants: the 4−point condition with log-det metric ([15], [50]), the studying of algebraic relationships among expected frequencies ([31], [32]), and harmonic analysis ([29]). In general, in the cited works, these techniques are restricted to very specific topological trees and models (for example, in [15], the authors produce invariants for the Jukes-Cantor 2-base model with 4 terminal taxa) with few hopes to extend them to the general case. We suggest to look at the introduction of [4] for a better reference about these techniques.

Definition 2.3.5. *Let $(T_1, \mathcal{M}_1)$ and $(T_2, \mathcal{M}_2)$ be two general Markov models with the same number of leaves and with respectively joint distribution tensors P_1 and P_2. If a polynomial f is such that $f(P_1) = 0$ (or $f(P_1) \approx 0$) but $f(P_2) \neq 0$ (or $f(P_2) \not\approx 0$) we say that f is* **topologically informative**.

Example 2.3.6. Consider the model of Cavender and Felsenstein in [15]. This is a symmetric model with $k = 2$ states given by 0 and 1.



After the stochastic invariant, we have the linear invariants given by the symmetry of the model: $p_{0000} - p_{1111}$, $p_{0011} - p_{1100}$, $p_{0001} - p_{1110}$, $p_{0110} - p_{1001}$, etc. Finally, the last one is the **informative** invariant:

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

The origin of this invariant can be found in the 4−point condition for tree metrics. This polynomial vanishes only for the 4-leaf tree where a and b are neighbours, and does not vanish for generic joint distributions arising from the other two 4-leaf topologies (given by the other two ways to label leaves, with a in the same cherry of c or d). Thus f is topologically informative.

2.4 Phylogenetic ideal and phylogenetic variety

As already said, a model $\mathcal{M}$ on a tree T, with n leaves, has $N := (k-1) + (2n-3)k(k-1)$ free parameters π_i's and $m_{ij}^{(e)}$'s. Thus, the stochastic parameter space for the tree T is given by $S \subset [0,1]^N$, and each $s \in S$

represents a model $\mathcal{M} = (\pi(r), \{M^{(e)}\})$. Using Formula (2.24), we can define a parameterization map

$$\begin{array}{rccc} \varphi_T : & S & \to & [0,1]^{k^n} \\ & s & \to & P = [p_{11\cdots 1}, \ldots, p_{kk\cdots k}] \end{array} \tag{2.26}$$

where $[0,1]^{k^n}$ represents the joint distribution state. An element $P \in [0,1]^{k^n}$ in the image of φ_T represents a joint distribution of pattern frequencies at the leaves of T.

Since, by Formula (2.24), φ_T is a polynomial map in the unknown parameters, $(T, \mathcal{M})$ is a parametric algebraic statistical model (Definition 1.1.8). Thus, we can use the same approach of Section 1.5 to define ad ideal and a variety associated to $(T, \mathcal{M})$. As a matter of fact, also in this case, we can extend the parameterization to

$$\Phi_T : \mathbb{C}^N \to \mathbb{C}^{k^n}. \tag{2.27}$$

Definition 2.4.1. *Let V_T be the (Zarisky) closure of the image of Φ_T, that is $V_T := \overline{\Phi_T(\mathbb{C}^N)}$. V_T is called the* **phylogenetic variety** *associated to the tree T. The ideal $I_T \subset \mathbb{C}[p_{i_1\cdots i_n}]$ of V_T is called the* **phylogenetic ideal**.

Remark 2.4.2. In literature we can find a different (but equivalent) definition of phylogenetic ideal. In that case the phylogenetic ideal I_T is define as the ideal generated by the phylogenetic invariants of the general Markov model $(T, \mathcal{M})$. This follow from the fact that if two polynomials of $\mathbb{C}[p_{i_1\cdots i_n}]$ vanish under such substitution, then so do any of their linear combinations with coefficients in $\mathbb{C}[z_{i_1\cdots i_n}]$. From Chapter 0 of [26], it follows that the phylogenetic invariants form an ideal in R. Successively, from the definition of the phylogenetic ideal I_T we obtain the phylogenetic variety as the variety associated to I_T.

Roughly speaking, V_T is a variety that contains the (complex) joint distribution for all possible choices of (complex) numerical parameters $\mathcal{M} = (\pi(r), \{M^{(e)}\})$ of general Markov model on the tree T. In applications, the tree topology is usually the parameter of greatest interest. If an observed distribution of pattern frequencies were "close" to V_T, that could be interpreted as support for inferring T. Thus, phylogenetic invariants allow the inference of T without having to estimate all the other parameters, as, on the contrary, maximum likelihood requires.

Remark 2.4.3. Extending the parameterization Φ_T to the complex numbers from the stochastic setting is done because an algebraically closed field provides the easiest and most natural setting for understanding polynomial maps. Of course, complex parameters and complex joint distributions are not so natural from a biological or statistical viewpoint. Obviously, the final goal would be to understand the model in the stochastic setting.

The Algebraic Algorithm of Section 1.5 can be modified for the precise purpouse of general Markov models on trees. Once the number n of species is fixed, the algorithm will run over all possible trees T with n leaves.

Algebraic Algorithm for Phylogenetic Inference

INPUT: the joint distribution of observed frequencies $\hat{P}$.

i) fix M;

ii) for each tree T

- Find some/most/all invariants f for V_T;
- Test if $f(\hat{P}) \approx 0$.

OUTPUT: the tree T for which $\hat{P}$ is as close as possible to V_T.

Understanding V_T well means both theoretical and practical understanding of problems of phylogenetic inference. One part of understanding V_T is describing it implicitly, as the zero set of polynomials. This means to find polynomials $f \in R$ such that $f(q) = 0$ for all $q \in V_T$. This is equivalent to find the kernel of the map

$$\tilde{\Phi}_T : \mathbb{C}[z_{i_1 \cdots i_n}] \to \mathbb{C}[s_1, \ldots, s_N]$$

where $\mathbb{C}[s_1, \ldots, s_N]$ is the polynomial ring associated to vector space $\mathbb{C}^N$ (we can take, as variables s_i's, for example, the unknown stochastic parameters of the Markov model, $\pi(r)_l$, $m_{ij}^{(e)}$). The kernel of $\tilde{\Phi}_T$ is the ideal I of the polynomials in the $z_{i_1 \cdots i_n}$'s vanishing for all choices of (stochastic or complex) parameters s_i's, i.e. it corresponds exactly to the phylogenetic ideal as defined in Definition 2.4.1. As already said at the end of Section 1.5 to find explicitly the ideal of a general Markov model for each topological tree is equivalent to giving a list of generators of such ideal. In the specific case of phylogenetic ideals Gröbner basis and Elimination Theory show their limits. As a matter of fact the basic alghorithm to give I_T is the application of the elimination process to the set of equations $z_{i_1 \cdots i_n} - p_{i_1 \cdots i_n} = 0$ with respect to the indeterminates/parameters $\pi(r)_l$'s, $m_{ij}^{(e)}$'s. This process involves k^n polynomials of degree $2n - 3$ in the variables $\pi(r)_l$'s, $m_{ij}^{(e)}$'s. Thus, as soon as the number of leaves, or the number of states grows, the computation becomes more and more complex.

Remark 2.4.4. In finding phylogenetic invariants, the main goal is to determine the full ideal I_T, i.e. an ideal-theoretic definition of V_T. However, a weaker goal is to determine a set of polynomials whose zero

set is V_T. Namely, what we are looking for is a set-theoretic definition of the variety without determining the whole ideal.

Researchers, in the field of Phylogenetic Algebraic Geometry, are looking for different techniques and tools to find phylogenetic invariants.

In [4], Allman and Rhodes present several methods of finding phylogenetic invariants for the general Markov model of base substitution along any topological tree. In particular, the authors do not require any conditions on the numbers n and k of leaves and states. The constructions are based on commutation and symmetry relations of matrix expressions and that requires only linear algebra. In particular, for a 3-taxon tree T, a *strong set* of invariants can be found. These invariants have degree $k+1$ (the lowest possible) and have many terms. For example, if $k = 4$, they are all degree five invariants (1728−dimensional space) and the number of terms in each invariant is about 180. The constructions are successively generalized to the n−taxon case. Unluckily, here, these invariants do not generate the full ideal, but in some cases they give a satisfactory result. It is important to observe that, since they are expressed in matrix form, invariants may be evaluated through numerical linear algebra.

A fully observable homogeneous Markov model has no hidden nodes and all matrices are the same. An explicit analysis of this kind of model, on a tree with at most five leaves and $k = 2$, can be found in [27].

Remark 2.4.5. Although we consider here only the part of the study of phylogenetic invariants concerned with Algebraic Geometry, it is mandatory to recall that a statistical understanding of the behaviour of these polynomials is necessary. This is due, in particular, to the fact that we want to apply invariants to real and noisy data. Moreover, the models of evolution are only an approximation of reality, then, from a statistical point of view, we need the robustness of method under violation of model assumptions. Finally, Statistics will be necessary for a good defintion of "close to vanish".

We now fix our attention on the technique of flattening of a n−dimensional tensor. As we will see in the next two Sections, the flattenings permit to obtain invariants from the "local" structure of the tree. Moreover, the flattenings are strictly connected with the theory of secant varieties. There is a huge literature on this theory; some papers are becoming fundamental in the research of phylogenetic invariants.

2.5 Flattenings

We introduce, now, the notion of **edge flattening** of a tensor $P \in \mathbb{C}^{k^n}$ according to an n−taxon tree T. Let $P = \Phi_T(s)$ be the joint distribution of states for some parameters choice for the general Markov model on T. Consider an edge e of T. Then e induces a split of the taxa according to the connected components of

$T \setminus \{e\}$. We can assume, eventually re-ordering the indices in P, that the split is $\{\{l_1, \ldots, l_t\}, \{l_{t+1}, \ldots, l_n\}\}$. We can imagine now a statistical model based on the split induced by e: we can group the taxa $\{l_1, \ldots, l_t\}$ and the taxa $\{l_{t+1}, \ldots, l_n\}$, so that each is on a leaf attached to a common ancestral node, which is choosen to be on one vertex of e (Fig. 6). The joint states at the taxa $\{l_1, \ldots, l_t\}$, are described through a single k^t-state variable and similarly at $\{l_{t+1}, \ldots, l_n\}$ are described through a single $k^{n-t}-$state variable. Thus, we have a coarser graphical model with one hidden $k-$state internal node and two descendant nodes with respectively k^t and k^{n-t} states. Forming the joint distribution for this coarser model, we get a $k^t \times k^{n-t}$ matrix $Flat_e(P)$ which is defined in the following way: fix any ordering of $J_1 := [k]^t$ and $J_2 := [k]^{n-t}$ and for $u \in J_1$, $v \in J_2$ let

$$Flat_e(P)(u, v) = P(u_1, \ldots, u_t, v_1, \ldots, v_{n-t})$$

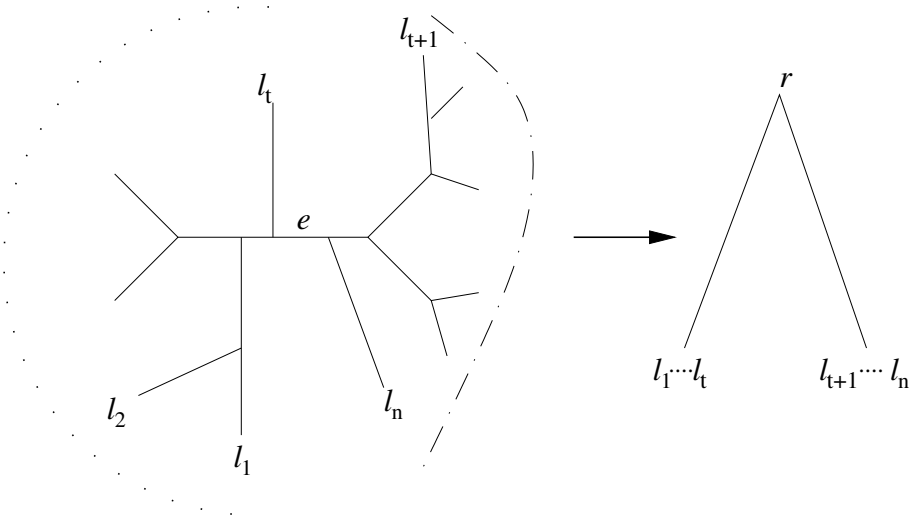
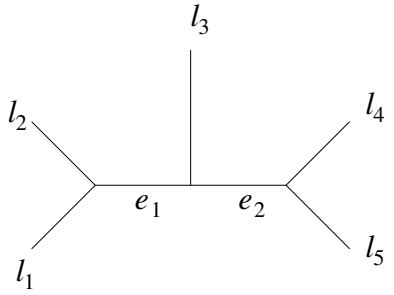


Fig. 6: An example of edge flattening

Then $Flat_e(P)$ can be seen as a joint distribution for a related graphical model with a less complicated structure: one hidden $k-$state internal node and two descendant nodes with k^t and k^{n-t} states, respectively.

Example 2.5.1. Consider the following $5-$leaf tree T



with $k = 2$ states at each vertex, represented by 0 and 1. The two splits

$$\{\{l_1, l_2\}, \{l_3, l_4, l_5\}\} \text{ and } \{\{l_1, l_2, l_3\}, \{l_4, l_5\}\}$$

give respectively the flattenings

$$Flat_{e_1}(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix} \tag{2.28}$$

and

$$Flat_{e_2}(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix} \tag{2.29}$$

Here, for example, the $(01, 000)-$entry of $Flat_{e_1}(P)$ is the probability of observing state 01 at leaf $\{l_1, l_2\}$, and state 000 at leaf $\{l_3, l_4, l_5\}$. Of course, this entry is precisely p_{01000}.

Remark 2.5.2. A combinatorial result, the Splits Equivalence Theorem, states that a tree is uniquely determined by its set of splits. See [SS] for a proof.

For the coarser graphical model, the joint distribution matrix must have the form

$$Flat_e(P) = M_1^T \, diag(\pi(r)) M_2 \tag{2.30}$$

where M_1 and M_2 are $k \times k^t$ and $k \times k^{n-t}$ matrices and $diag(\pi(r))$ is a diagonal matrix with $(i, i)-$entry $\pi(r)_i$ (the precise description of M_1 and M_2 can be found in [7]).

Remark 2.5.3. The coarser model is not a phylogenetic tree since the number of states at the two leaves are different powers of k.

From (2.30) we quickly obtain that $rank(Flat_e(P)) \leq k$. Hence, all $(k+1)\times(k+1)$ minors of $Flat_e(P)$ must vanish. These minors generate the full ideal of polynomials vanishing on matrices of rank $\leq k$, and thus we have found all phylogenetic invariants associated to the coarser model. Such invariants are also invariants for the original model on T, and they are called **edge invariants** associated to the edge e. Moreover, we denote by $\mathcal{F}_{edge}(T)$ the set of all $(k+1)\times(k+1)$ minors of the edge flattenings $Flat_e(P)$ as e varies on E. An important result concerning flattenings is the following

Theorem 2.5.4 (E. Allman, J. Rhodes, [6]). *For $k = 2$ and any number of taxa n, the phylogenetic ideal I_T, for the general Markov model $\mathcal{M}$ on an n-taxon tree T, is generated by $\mathcal{F}_{edge}(T)$, the 3×3 minors of all edge flattenings of a $2\times 2\times\cdots\times 2$ tensor of indeterminants.*

Thus, in particular one has

Corollary 2.5.5. *For the 5-leaf tree T of Example 2.5.1, I_T is generated by all 3×3 minors of matrices (2.28) and (2.29).*

However, for a larger k, it is not enough to consider only 2−dimensional edge flattenings (i.e., flattenings to matrices) to obtain generators of the phylogenetic ideal. Consider, for example, a 3−taxon tree T: if $k > 2$ we know that I_T contains polynomials of degree $k+1$ although $\mathcal{F}_{edge}(T)$ is empty. Hovewer, we can give an interesting result.

Proposition 2.5.6 (E. Allman, J. Rhodes, [6]). *For any k−state general Markov model on T, or any submodel, the phylogenetic ideal contains all edge invariants.*

To find other invariants, we can introduce another kind of flattening which produces 3−dimensional tensors. Consider an internal node v of T, contained in edges e_1,e_2,e_3. Then v induces a tripartition of the taxa according to the connected components of $T\setminus\{v, e_1, e_2, e_3\}$. We may assume the tripartition is

$$\{\{l_1,\ldots,l_{n_1}\},\{l_{n_1+1},\ldots,l_{n_1+n_2}\},\{l_{n_1+n_2+1},\ldots,l_{n_1+n_2+n_3}\}\}.$$

where $n_1+n_2+n_3 = n$. Then a vertex flattening of P at v is a $(k^{n_1}\times k^{n_2}\times k^{n_3})$−array $Flat_v(P)$ defined as follows: fix an ordering of $J_1 = [k]^{n_1}$, $J_2 = [k]^{n_2}$, $J_3 = [k]^{n_3}$, and for $x\in J_1$, $y\in J_2$, $z\in J_3$ let

$$Flat_v(x,y,z) = P(x_1,\ldots,x_{n_1},y_1,\ldots,y_{n_2},z_1,\ldots,z_{n_3}).$$

Thus, the final result of a vertex flattening is a graphical model with one hidden k−state internal node and three descendant nodes with k^{n_1}, k^{n_2}, and k^{n_3} states, respectively. Since an ideal is associated to such a graphical model, we can talk of the ideal of the vertex flattening.

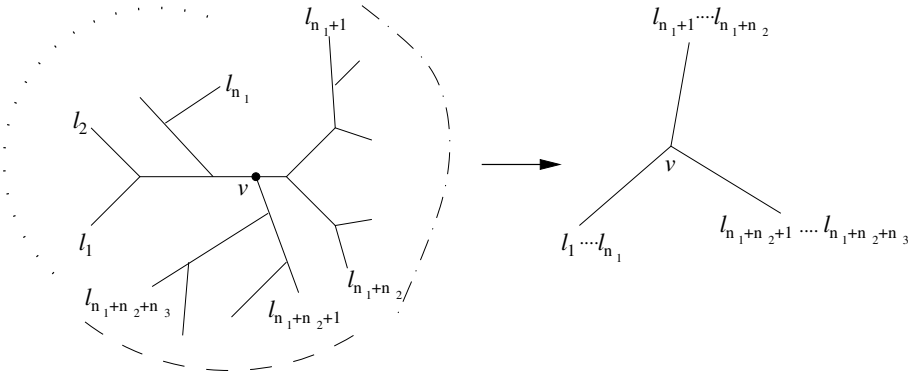


Fig. 7: An example of vertex flattening

We have now all the elements to give the following

Conjecture 2.5.7 (E. Allman, J. Rhodes, [6])**.** *For any k and any number of taxa n, the phylogenetic ideal I_T, for the general Markov model on an $n-$taxon tree T, is the sum of the ideals associated to the flattenings of P at vertices of T.*

It is important to remark that this Conjecture, for $k = 2$, is identical to Theorem 2.5.4. In fact, by the results of Lansdberg and Manivel ([44]) we know that, in this case, the ideal associated to a vertex flattening is the sum of the ideals associated to the edge flattenings of the three edges containing the vertex.

2.6 Secant varieties

In the previous section we only considered matrices with rows that sum to 1. This probabilistic condition can be interpreted in Algebraic Geometry as the fact that each row of a transition matrix is an element of a certain affine subspace of a projective space $\mathbb{P}^{k-1}$. At the same time, as in Lecture 1, we can view V_T projectively. In fact, by the stochastic invariant, one has $V_T \subset \mathbb{P}^{k^n-1}$. The passage to the projective case forces us to look only for phylogenetic invariants among homogeneous polynomials.

Consider a 3-taxon rooted star tree T in the projective setting for k states.

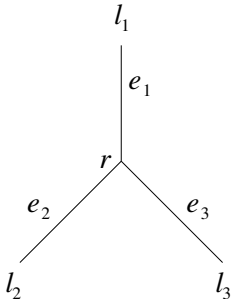


Fig. 8: The $3-$taxon tree T_3

Using the same construction of Section 1.6 we know that V_T is $S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \mathbb{P}^{k-1})$, i.e. the $k-$secant variety of the Segre product of three $\mathbb{P}^{k-1}$.

We have to point out that the root distribution does not explicitly appear, since it has been accounted for in the arbitrary scaling factors that appear in each P^i, when we choose particular projective coordinates to express them. Hence, the joint distribution P has been decomposed as the sum of k rank 1 tensors, one for each possible state at the root (and this forces P to have rank k, by the definition itself of tensor rank).

Example 2.6.1. Consider $k = 2$. The general Markov model on T_3 has only 7 parameters and since the stochastic invariant cuts out a $7-$dimensional subspace of $\mathbb{C}^{2^3}$, one could expect that there are no other invariants. In fact, one has $S_2(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1) = \mathbb{P}^7$ (i.e. every $2 \times 2 \times 2$ tensor is in the closure of the rank 2 tensors). We want to underline that for the 3-taxon tree, the construction of edge invariants yields nothing, since there are no internal edges.

Example 2.6.2. Go on considering T_3, but with $k = 3$. In this case the ideal defining $S_3(\mathbb{P}^2 \times \mathbb{P}^2 \times \mathbb{P}^2)$ was found in [33], and given in terms of Bayes models. Let $A = (p_{ij1})$, $B = (p_{ij2})$ and $C = (p_{ij3})$ be three $3 \times 3-$matrices obtained by taking slices of the $3-$dimensional tensor P associated to the model. Then one has

Proposition 2.6.3 (L. D. Garcia, M. Stillman, B. Sturmfels, [33])**.** *Let I be the ideal of $Sec^{k-1}(\mathbb{P}^2 \times \mathbb{P}^2 \times \mathbb{P}^2)$, the naive Bayes model with $n = 3$ ternary features with k classes. If $k = 2$ then I is generated by the $3 \times 3-$subdeterminantal of any two-dimensional table obtained by flattening the $3-$dimensional tensor P. If $k = 3$ then I is generated by the quartic entries of the various $3 \times 3-$matrices of the form*

$$A \cdot adj(B) \cdot C - C \cdot adj(B) \cdot A.$$

If $k = 4$ then I is the principal ideal generated by the following homogeneous polynomial of degree 9 with 9, 216 terms:

$$det(B)^2 \cdot det(A \cdot B^{-1} \cdot C - C \cdot B^{-1} \cdot B).$$

If $k \geq 5$ then I is the zero ideal.

More generally, let T be a star tree with an internal node r and n leaves l_i. We can suppose that the hidden variable associated to r has k states, while the hidden variable at the leaf l_i has a_i states, $i = 1, \ldots, n$. The variety associated to this model is the $k-$secant variety of the Segre product $\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1}$.

There is a useful relationship between the varieties $S_k(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$ and $S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{k-1})$. In fact, for any joint distribution $P \in S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{k-1})$, there is an "action" by $k \times a_n$ complex matrices M in the last index of P. This gives us a point $P *_n M \in S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{a_n-1})$

(here k is repeated $n-1$ times). Using the map Φ_T, if $P = \Phi_T(\pi(r), \{M_1, M_2, \ldots, M_n\})$, where M_n is the matrix on the edge leading to the $n-$th leaf, then $P *_n M = \Phi_T(\pi(r), \{M_1, M_2, \ldots, M_{n-1}, M_n M\})$, though the action extends to the points (on the variety) that are not in the image of Φ_T. We can define also an "action" by $a_n \times k$ matrices N on $S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{a_n-1})$. Thus, given a $k \times a_n$ matrix M and an $a_n \times k$ matrix N we have maps

$$S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{k-1}) \underset{*_n N}{\overset{*_n M}{\rightleftarrows}} S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{a_n-1}) \tag{2.31}$$

From these maps, we can obtain maps between the ideals of the two varieties. The compositions of these maps are related to $GL(k, \mathbb{C})$ and $GL(a_n, \mathbb{C})$ actions. In a similar way, we can define an action on each distinct index, not just on the last one. Thus, we obtain an action of $GL(a_1, \mathbb{C}) \times GL(a_2, \mathbb{C}) \times \cdots \times GL(a_n, \mathbb{C})$ on $S_k(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$. We have the following

Theorem 2.6.4 (E. Allman, J. Rhodes, [6]). *Suppose $a_1, a_2, \ldots, a_n \geq k$. Let $\mathcal{F}$ be any set of polynomials whose zero set is $S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{k-1})$. For $t = 1, 2, \ldots, n$, let $Z_t = (z^t_{ij})$ be $a_t \times k$ matrices of indeterminants. For an $(a_1 \times a_2 \times \cdots \times a_n)-$ tensor P of indeterminants, let $\tilde{P}$ be the $(k \times k \times \cdots \times k)-$tensor that results from letting each Z_t acts formally in the $t-$th index of P (i.e. as the lower map in (2.31)). Let $\tilde{\mathcal{F}}$ denote the set of polynomials in the entries of P obtained from those in $\mathcal{F}$ by substituting into them the entries of P, expressing the results as polynomials in the z^t_{ij}, and then extracting the coefficients. Let $\mathcal{F}_{edge}$ be the set of $(k+1) \times (k+1)$ minors of the n flattenings of P on edges of the star tree. Finally, let $\mathcal{F}(k; a_1, a_2, \ldots, a_n) = \tilde{\mathcal{F}} \cup \mathcal{F}_{edge}$. Then $\mathcal{F}(k; a_1, a_2, \ldots, a_n)$ defines $S_k(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$ set-theoretically.*

Similarly, an ideal-theoretic version of this result can be given:

Theorem 2.6.5 (E. Allman, J. Rhodes, [6]). *Suppose $a_1, a_2, \ldots, a_n \geq k$. Let $\mathcal{F}$ be any set of polynomials generating the ideal of $S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \cdots \times \mathbb{P}^{k-1})$. Then the set $\mathcal{F}(k; a_1, a_2, \ldots, a_n)$ generates the ideal of $S_k(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$.*

Remark 2.6.6. Although Allman and Rhodes define $S_k(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$ as the variety associated to a star tree with t leaves, we obviously have to consider only $n = 3$, because the tree is bifurcating. We have to point out that a set of polynomials, defining the variety $S_k(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$, when $k = 2$, can be found in [44].

Consider a vertex flattening on a tree T. Now, the variety associated to the coarsened model is the variety of rank k tensors of size $k^{n_1} \times k^{n_2} \times k^{n_3}$, i.e. $S_k(\mathbb{P}^{k^{n_1}-1} \times \mathbb{P}^{k^{n_2}-1} \times \mathbb{P}^{k^{n_3}-1})$. It is important to observe

that, in such a case, one has $a_i = k^{n_i}$ for an integer n_i depending on the splitting. Thus, the hypotheses $a_i \geq k$ are satisfied and, in some way, we can try to use polynomials vanishing on $S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \mathbb{P}^{k-1})$ to obtain invariants for the vertex flattening $S_k(\mathbb{P}^{k^{n_1}-1} \times \mathbb{P}^{k^{n_2}-1} \times \mathbb{P}^{k^{n_3}-1})$ and then for the starting tree T. More precisely, one has the following

Theorem 2.6.7 (E. Allman, J. Rhodes, [6])**.** *For a* 3*−leaf star tree, let* $\mathcal{F}$ *be a set of polynomials defining* $S_k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \mathbb{P}^{k-1})$ *set-theoretically, and let* $\mathcal{F}(k; a_1, a_2, a_3)$ *be as defined in Theorem 2.6.4. For an* n*−taxon tree* T*, let* $\mathcal{F}(T)$ *be the union of all sets* $\mathcal{F}(k; k^{n_1}, k^{n_2}, k^{n_3})$ *associated to* 3*−dimensional flattenings at nodes of* T*. Then the zero set of* $\mathcal{F}(T)$ *is the phylogenetic variety* V_T*.*

In several cases, edge flattenings and vertex flattenings permit to determine, at least set-theoretically, the phylogenetic variety. We can then investigate if different kinds of flattenings can give new phylogenetic invariants (see [6], page 12). In any case, the previous theorems seem to suggest that the phylogenetic variety is determined by the local structure of the tree and encourages Phylogenetic Algebraic Geometry in this direction.

We can conclude with a remark about Theorem 2.6.4. An important consequence of this Theorem is the following

Corollary 2.6.8. *For* $n \leq 5$*, the ideal of the variety* $S_2(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$ *associated to the hidden naive Bayes model with a* 2*−state hidden variable and* n *observed variables with* $a_1, \ldots, a_n$ *states, is generated by the* 3×3 *minors of all* 2*−dimensional flattenings associated to bipartitions of the observed variables.*

We have to point out that the case $n = 3$ was already proved in [44]. The previous Corollary solves several cases of the following

Conjecture 2.6.9 (L. D. Garcia, M. Stillman, B. Sturmfels, [33])**.** *The prime ideal* Q_G *of any naive Bayes model* G *with* $r = 2$ *classes is generated by the* 3×3*−subdeterminants of any 2-dimensional table obtained by flattening the* n*−dimen-sional table* $(p_{i_1 i_2 \cdots i_n})$*.*

Theorem 2.6.4 limits the study of the ideal of $S_2(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$ to the "simplest" case $S_2(\mathbb{P}^1 \times \mathbb{P}^1 \times \cdots \times \mathbb{P}^1)$, since, applying the construction of this Theorem and maps as in (2.31), we obtain the generators for the variety $S_2(\mathbb{P}^{a_1-1} \times \mathbb{P}^{a_2-1} \times \cdots \times \mathbb{P}^{a_n-1})$, with $a_i \geq 2$. Thus, the Conjecture can be restated as

Conjecture 2.6.10 (L. D. Garcia, M. Stillman, B. Sturmfels, [33])**.** *The ideal of the variety* $S_2(\mathbb{P}^1 \times \mathbb{P}^1 \times \cdots \times \mathbb{P}^1)$*, that is, the ideal associated to the hidden naive Bayes model with a* 2*−state variable and* n

2−*state observed variables, is generated by the* 3×3*-subdeterminants of all two-dimensional flattenings arising from bipartitions of the observed variables.*

Lecture 3. Identifiability, Bernoulli models, decomposition of tensors

3.1 Identifiability

Consider a parametric algebraic statistical model $\mathcal{M}_f$ which is described by a parameterization map

$$\begin{array}{rccc} f: & \Theta \subseteq \mathbb{R}^d & \to & \mathbb{R}^m \\ & \theta & \mapsto & (f_1(\theta), \ldots, f_m(\theta)) \end{array}$$

Definition 3.1.1. *The model $\mathcal{M}_f$ is* **(strict) identifiable** *if f is injective.*

The identifiability of a model is a foundamental property in Statistics, since inference will produce an unique result on such model.

It is important to notice that in many cases, the above map will not be strictly injective. For example, when we work with a general Markov model on tree, the above map is always $r!$-to-one, where r is the number of hidden classes in the model (label-swapping effect). However a finite number of solution is still better than an infinite number of solutions !!!

The concept of identifiability in Statistics has an analogue in Projective Geometry. We use now definitions and results given in Appendix, Section A.2.5.

We work over the complex field and we consider the projective space $\mathbb{P}^r = \mathbb{P}^r_{\mathbb{C}}$, equipped with the tautological line bundle $\mathcal{O}_{\mathbb{P}^r}(1)$.

Definition 3.1.2. *A projective variety $X \subset \mathbb{P}^r$ is called* **k-identifiable** *if the general element of $S_k(X)$ has a unique decomposition as the sum of k elements of X.*

An equivalent definition can be given in term of secant order of X (see Definition A.2.11)

Definition 3.1.3. *We say that X is* **(generically) k-identifiable** *if it has k-th secant order 1, i.e. if for a general point $P \in S_k(X)$, there is a unique unordered k-uple $P_1, \dots, P_k$ of points of X such that $P \in \langle P_1, \dots, P_k \rangle$.*

It is clear that X is k-identifiable when the projection $AbS_k(X) \to S_k(X)$ is birational. For this reason X cannot be k-identifiable when $\dim(AbS_k(X)) > S_{(k)}(X)$. In particular, if X has dimension m then X is not k-identifiable when $r < k(m+1)$ or when X is k-defective.

The main link between identifiability and weakly defective varieties lies in the following:

Theorem 3.1.4 (L. Chiantini, C. Ciliberto, [18], Corollary 2.7)**.** *Let $X \subset \mathbb{P}^r$ be an irreducible, projective, non–degenerate variety of dimension m. Assume $k(m+1) - 1 < r$. Then X is k-identifiable, unless it is k–weakly defective.*

Theorem 3.1.5 (L. Chiantini, C. Ciliberto, [18], Theorem 2.4)**.** *Let $X \subset \mathbb{P}^r$ be an irreducible, projective, non–degenerate variety of dimension m. Assume $k(m+1) - 1 < r$ and assume that X is k-weakly defective. Call Σ a general k-th contact variety. Then, the k-th secant order of Σ is equal to the k-th secant order of X.*

Thus, a way to prove that a variety X is k-identifiable, at least when $r \neq k(m+1)-1$, is to prove that X is not k-weakly defective, or, if it is k-weakly defective, that the general contact variety Σ has k-th secant order 1.

Using the connection between k-defectivity and k-weakly defectivity in term of the existence of degenerate subvarieties, passing through k general points in X, we can use Corollary A.2.17 for identifiability.

In conclusion, we obtain:

Corollary 3.1.6. *Assume $r > k(m+1) - 1$. Assume that for all $n = 1, \dots, m-1$, there are no families of n-dimensional subvarieties of X, whose general element spans a linear space of dimension $\leq k(n+1) - 1$ and passes through $k+1$ general points of X. Then X is not k-weakly defective. Hence it is k-identifiable.*

One should observe that both Corollary 3.1.6 and the second part of Theorem 3.1.4 cannot be inverted.

Example 3.1.7. When X is k-weakly defective, it can be k-identifiable as well. This may happen, by [18], Theorem 2.4, when the contact locus has k-th secant order 1. Examples of such varieties can be found in [18], Example 3.7, but they are singular. A smooth example was communicated us by G. Ottaviani. Take the Segre embedding of $X = \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^2$ in $\mathbb{P}^{11}$. Using a computer-aided procedure, one can find that the general hyperplane which is tangent to X at two points, is indeed tangent along a twisted cubic. The computation was indeed performed at two specific points of X, but notice that $\mathrm{Aut}(X)$ acts transitively on pair of points. Thus X is 1-weakly defective. Since a twisted cubic curve has first secant order equal to 1, it turns out by [18], Theorem 2.4, that X is 1-identifiable. The 1-identifiability of X also follows from the Kruskal's identifiability criterion for the product of three projective spaces (see [42]).

As a consequence, one cannot use the inverse of the previous argument to determine the non-identifiability of a variety X, simply by studying degenerate subvarieties.

Those who wonder why we ask for the uniqueness just for a *general* element of $S_k(X)$, should consider that for any $k \geq 2$ there are always points of $S_k(X)$ which have rank smaller than k.

Notice that k-identifiable implies $(k-1)$-identifiable and so on. In a similar way, $k-$weak-defectivity implies $(k-1)-$weak-defectivity and so on.

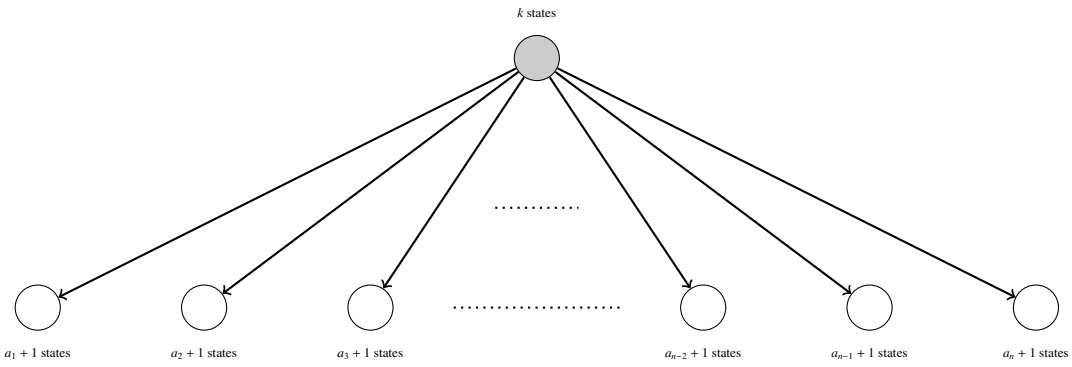
A fundamental Geometric tool for the analysis of the identifiability of tensors, is the following proposition, which is essentially a consequence of Terracini Lemma.

Proposition 3.1.8 (L. Chiantini, G. Ottaviani, [19], Proposition 2.4)**.** *If there exists a set of k particular points $x_1, \dots x_k \in X$, such that the span*

$$< \mathbb{T}_{x_1} X, \dots, \mathbb{T}_{x_k} X >$$

contains $\mathbb{T}_x X$ only if $x = x_i$ for some $i = 1, \dots k$, then X is k-identifiable.

Now we fix our attention on $k-$secant varieties of Segre products $X = \mathbb{P}^{a_1} \times \cdots \times \mathbb{P}^{a_n}$. By Section 1.6 we know that $S_k(X)$ is the variety associated to the following graphical model.



By definition, the $k-$identifiability of X implies the identifiability of the model.

If we do not want to think at $S_k(X)$ as the variety associated to a statistical model, we can still consider it as the set of $(a_1 \times a_2 \times \cdots \times a_n)-$tensors of rank $\leq k$. In this case the identifiability of X is strictly related to then uniqueness of decomposition of tensors as a sum of decomposable tensors. Conditions which guarantee the uniqueness of this decomposition are quite important in the applications [40]. Indeed, many decomposition algorithms converge to *one* decomposition, so that a uniqueness result guarantees that the decomposition found is the one we looked for. Even from a purely theoretical point of view, the study of

the decomposition shows some beautiful and not expected phenomena. After a look at the table in Section 3.7, we see that there are some exceptional sporadic cases which are intriguing.

So, one of the topics of this lecture is to show how to add the words identifiability/weak-defectivity to our Algebra/Geometry-Statistics Dictionary:

Statistics	Algebra
Independence	Segre Variety
Binomial Random Variable	Rational Normal Curve
Log-linear Model	Toric Variety
Mixture Model	Secant Variety
ML Estimation	Tropicalization
Design	Zero-dimensional Scheme
Identifiability	**Weak-Defectivity**
⋮	⋮

At the same time, our results on identifiability of models will produce also results on decompositions of tensors.

3.2 The main lemma

The inductive step, that allows us to provide effective results on the identifiability of tensors and statistical models on star graphs, relies in the following:

Lemma 3.2.1 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *Let X be a smooth non-degenerate projective subvariety of $\mathbb{P}^N$, of dimension n. Let Y denote the canonical Segre embedding of $X \times \mathbb{P}^m$ into $\mathbb{P}^M$, $M = mN + m + N$. Fix k with $(n+1)k < N+1$ and $r < N$ such that $r + 1 \geq (n+m+1)k$. Assume that a general linear subspace of $\mathbb{P}^N$, of dimension r, which is tangent to X at k general points, is not tangent to X elsewhere.*

Then the general linear subspace of $\mathbb{P}^M$, of dimension $mr + m + r$, which is tangent to Y at $(m+1)k$ general points, is not tangent to Y elsewhere.

Proof. First of all, notice that $\dim(Y) = (m+n)$, and $(m+1)(r+1) \geq (m+n+1)(m+1)k$. Thus, by an obvious parameter count, there are linear subspaces of dimension $mr + m + r$ which are tangent to Y at $(m+1)k$ general points.

Fix $m+1$ independent points $p_0, \dots, p_m$ of $\mathbb{P}^m$ and for $j = 0, \dots, m$ take k general points q_{ij} of the fiber $X \times \{p_j\}$. Call π_j the natural projection of $X \times \{p_j\}$ to X.

For $h = 0, \dots, m$, fix a general linear subspace R_h, of dimension r, which is tangent to $X \times \{p_h\}$ at the k points $q_{1h}, \dots, q_{kh}$ and passes through the points $\pi_j(q_{ij}) \times \{p_h\}$, for $j \neq h$. Since $r + 1 \geq k(n+1) + km$, such spaces R_h exist. Moreover R_h it is tangent to $X \times \{p_h\}$ only at the point $q_{1h}, \dots, q_{kh}$, by our assumption on X.

Let R be the span of all the R_h's. We claim that R, which is a linear subspace of dimension $mr + m + r$, is tangent to Y at all the points q_{ij}, and it is not tangent to Y elsewhere. This will conclude the proof of the lemma, by semicontinuity.

First notice that for all i, j, R contains $m+1$ general points of $\{\pi_j(q_{ij})\} \times \mathbb{P}^m$, hence it contains these fibers. Since R also contains the tangent spaces to $X \times \{p_h\}$ at the points q_{ih}'s for all h, then it is tangent to Y at all the points q_{ij}'s.

Assume now that there exists a point $x \in Y$, different from the q_{ih}'s, such that R is tangent to Y at x. Call x' the projection of x to $\mathbb{P}^m$, so that in some coordinate system, we can write $x' = a_0 p_0 + \cdots + a_m p_m$. There is at least one of the a_i's, say a_0, which is non-zero. Assume that also $a_1 \neq 0$. Then, the projection of R to $\mathbb{P}^N \times \{p_0\}$, which by construction coincides with R_0, is also tangent to $X \times \{p_0\}$ at the projection of q_{k1}. By the generality of the choice of the q_{ij}'s, q_{k1} cannot coincide with any of the points $q_{10}, \dots, q_{k0}$. Thus we get a contradiction.

So, we conclude that $a_1 = 0$. Similarly we get that $a_2 = \cdots = a_m = 0$. It follows that $x = x'$ belongs to $X \times \{p_0\}$ and since R_0 is tangent to $X \times \{p_0\}$ at x, then x must coincide with some point q_{i0}. □

Remark 3.2.2. It is worthy of spending one Remark to point out that, by semicontinuity, if a general linear subspace of $\mathbb{P}^N$, of dimension r, which is tangent to X at k general points, is not tangent to X elsewhere, then the same phenomenon occurs for general linear subspaces of dimension $r-1$, $r-2$, and so on.

The Lemma, together with Proposition 3.1.8, produces the following general principle:

Corollary 3.2.3. *With the same assumptions on X of Lemma 3.2.1, then $Y = X \times \mathbb{P}^m$ is $(m+1)k$-identifiable.*

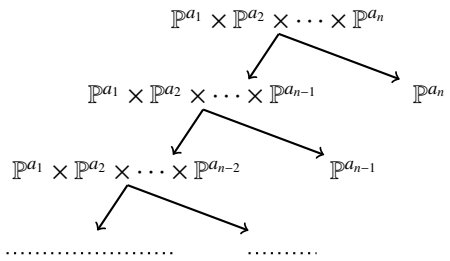
3.3 Results on Segre products

Let X be the Segre product $\mathbb{P}^{a_1} \times \cdots \times \mathbb{P}^{a_q}$ embedded in the usual way in $\mathbb{P}^N$ where $N + 1 = \prod_{i=1}^{q}(a_i + 1)$. It is well known that when k is bigger that the critical value

$$k_c := \frac{\prod_{i=1}^{q}(a_i + 1)}{1 + \sum_{i=1}^{q} a_i}$$

then the decomposition can never be unique (see the introduction of [11]). Only one case is known when k_c is an integer and there is a unique decomposition for tensors of rank equal to k_c, namely when $q = 3$ and $a_1 = 1, a_2 = a_3$.

So, let us consider the range $k < k_c$, where the problem can be understood better. Moreover, in this case, the Lemma 3.2.1 permits an inductive process, starting with a X who is a Segre product for which we know that the assumptions of Lemma 3.2.1 hold (by computer-aided specific computations or by Theorem 3.6.1 below) and then extending the number of factors of X.



3.4 Many copies of $\mathbb{P}^1$

The main case in which the previous result applies is the Segre product of many projective lines.

Proposition 3.4.1 (C. Bocci, L. Chiantini, G. Ottaviani, [13]). *Let X be the product of n copies of $\mathbb{P}^1$, $6 \leq n \leq 12$, naturally embedded in $\mathbb{P}^{2^n-1}$. Then for $k < k_c = \frac{2^n}{n+1}$ the linear span of k general tangent spaces at X, is not tangent to X elsewhere. In particular X is k-identifiable for all $k < k_c$.*

Proof. Just a computer-aided computation, following the algorithm presented in Section 3.8. In the case of 12 copies, the algorithm goes out of memory if implemented in a straightforward way. We used a "divide and conquer" technique to save memory, running in 2 hours on a PC with two processors at 2GHz. □

Theorem 3.4.2 (C. Bocci, L. Chiantini, G. Ottaviani, [13]). *For $n \geq 12$, let X be the product of n copies of $\mathbb{P}^1$, naturally embedded in $\mathbb{P}^N$, with $N = 2^n - 1$. Then for $r < 2^n - 2^{n-12}$ and for $k \leq \frac{r+1}{n+1}$, a general linear subspace of dimension r which is tangent to X at k points, is not tangent to X elsewhere.*

Proof. The proof goes by induction on $n \geq 12$. If $n = 12$, the claim follows from the previous Proposition and Remark 3.2.2.

Assume the claim holds for $n - 1$. Again by Remark 3.2.2, it suffices to prove the claim for $r + 1 = 2^n - 2^{n-12}$. Fix k as above. Notice that $k' := \lceil k/2 \rceil$ is at most $(2^{n-1} - 2^{n-13})/(n+1) + 1$, which is smaller than

$(2^{n-1}-2^{n-13})/n$ for $n \geq 12$. Thus we may apply induction: the general linear subspace of $P^{N'}$, $N' := 2^{n-1}-1$, of dimension $(2^{n-1}-2^{n-13})-1$, which is tangent to $X' := (\mathbb{P}^1)^{n-1}$ at k' points, is not tangent to X' elsewhere. The claim now follows directly from the Main Lemma 3.2.1. □

Remark 3.4.3. The assumption $n \geq 6$ is motivated by the fact that for 5 copies of $\mathbb{P}^1$, X is k-identifiable if and only if $k \leq 4$, while the general tensor of rank 5 has exactly two decompositions We recall that for 4 copies of $\mathbb{P}^1$, X is k-identifiable if and only if $k \leq 2$, while it is a result of Strassen that the general tensor of rank 3 has infinitely many decompositions.

For 3 copies of $\mathbb{P}^1$, X is k-identifiable if and only if $k \leq 2$, 2 being the general rank.

As a corollary, we get

Theorem 3.4.4 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *For $n \geq 12$, let X be the product of n copies of $\mathbb{P}^1$, naturally embedded in $\mathbb{P}^N$, with $N = 2^n - 1$. Then for $k \leq (2^n - 2^{n-12})/(n+1)$, X is k-identifiable.*

3.4.1 Results for Bernoulli models

Let we finish with a short account of the status of the art, for the identifiability of Bernoulli models, i.e. tensors in the span of $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$. After the paper of Strassen [51], and using methods of Algebraic Geometry, Elmore, Hall and Neeman proved in [37] the following asymptotic result: when the number n of factors is "very large" with respect to k, a, then the Segre product $\mathbb{P}^a \times \cdots \times \mathbb{P}^a$ is k-identifiable. A much more precise bound for identifiability of binary products was obtained by Allman, Matias and Rhodes. In [3] (Corollary 5) they proved that the product of n copies of $\mathbb{P}^1$ is k-identifiable when $n > 2\lceil \log_2(k+1) \rceil + 1$. Successively, using Geometric methods as well as a result by Catalisano, Gimigliano and Geramita ([14]), Bocci and Chiantini, in [11] improved the bound, showing that a product of $n > 5$ copies of $\mathbb{P}^1$ is k-identifiable for all k such that $k+1 \leq 2^{n-1}/n$. The case of 5 copies of $\mathbb{P}^1$ was shown to be exceptional. The bound, which happened to be the best known up to now, is substantially improved by Theorem 3.4.4. Let us compare the results of Allman-Matias-Rhodes and Bocci-Chiantini with the ones of Theorem 3.4.4 for some values of n with respect also to the critical value k_c.

n	k_c	AMR	BC	BCO
6	9	5	5	9
10	92	22	50	92
20	49932	724	26214	49920

The results of Allman, Matias and Rhodes gave a lower bound for 2^m which is quadratic with respect

to $k+1$. The main result of Bocci-Chiantini, in [11], proves that X is k-identifiable for $k < 2^{n-1}/n$, which is a little better than half way from the critical (maximal) value k_c. Finally Theorem 3.4.4 shows that X is k-identifiable for

$$k \leq \frac{2^n}{n+1}(1-\frac{1}{2^{12}}) = \frac{4095}{4096}\,\frac{2^n}{n+1} = 0,9997\ldots \cdot k_c$$

a sensible improvement, as n grows.

3.5 Many copies of $\mathbb{P}^2$ and $\mathbb{P}^3$

Let us see what happens with the Segre product of many projective planes.

Proposition 3.5.1 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *Let X be the product of n copies of $\mathbb{P}^2$, $4 \leq n \leq 6$, naturally embedded in $\mathbb{P}^{3^n-1}$. Then for $k < k_c = \frac{2^n}{n+1}$ the linear span of k general tangent spaces at X, is not tangent to X elsewhere. In particular X is k-identifiable for all $k < k_c$.*

Proof. Just a computer-aided computation, following the algorithm presented in Section 3.8. □

Remark 3.5.2. The assumption $n \geq 4$ is motivated by the fact that for 3 copies of $\mathbb{P}^2$, X is k-identifiable if and only if $k \leq 3$, while it is a result of Strassen ([51] §4) that the general tensor of rank 4 has infinitely many decompositions.

Theorem 3.5.3 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *For $n \geq 6$, let X be the product of n copies of $\mathbb{P}^2$, naturally embedded in $\mathbb{P}^N$, with $N = 3^n - 1$. Then for $r < 3^n - 3^{n-6}$ and for $k \leq (r+1)/(2n+1)$, a general linear subspace of dimension r which is tangent to X at k points, is not tangent to X elsewhere.*

Proof. The proof goes by induction on $n \geq 6$. If $n = 6$, the claim follows from the previous proposition and Remark 3.2.2.

Assume $n \geq 7$ and the claim holds for $n-1$. Again by Remark 3.2.2, it suffices to prove the claim for $r+1 = 3^n - 3^{n-6}$. Fix k as above. Notice that $k' := \lceil k/3 \rceil$ is at most $(3^{n-1} - 3^{n-7})/(2n+1) + 1$, which is smaller than $(3^{n-1} - 3^{n-7})/(2n-1)$ for $n \geq 5$. Thus we may apply induction: the general linear subspace of $P^{N'}$, $N' := 3^{n-1} - 1$, of dimension $(3^{n-1} - 3^{n-7}) - 1$, which is tangent to $X' := (\mathbb{P}^2)^{n-1}$ at k' points, is not tangent to X' elsewhere. The claim now follows directly from the Main Lemma 3.2.1. □

As a corollary, we get

Theorem 3.5.4 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *For $n \geq 6$, let X be the product of n copies of $\mathbb{P}^2$, naturally embedded in $\mathbb{P}^N$, with $N = 3^n - 1$. Then for $k \leq (3^n - 3^{n-6})/(2n+1)$, X is k-identifiable.*

The previous result shows that X is k-identifiable for

$$k \leq \frac{3^n}{2n+1}(1-\frac{1}{3^6}) = \frac{728}{729}\,\frac{3^n}{2n+1},$$

i.e. up to $728/729 = 0.998\ldots$ of the critical (maximal) value k_c.

And now the reader can see how the trick goes, at least for cubic tensors. Once one determines a starting point, for few copies of given projective spaces (e.g. by using a computer-aided computation), then the Main Lemma 3.2.1 provides an extension to the product of an arbitrary number of copies of projective spaces, in which the bound is expressed as a constant fraction of the critical value k_c.

We end the list of particular cases with the product of many copies of $\mathbb{P}^3$, which is relevant because of its connection with the Algebraic Statistics of DNA chains.

Theorem 3.5.5 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *Let X be the product of $n \geq 5$ copies of $\mathbb{P}^3$, naturally embedded in $\mathbb{P}^N$, with $N = 4^n - 1$.*

(i) for $n = 5$, a general linear subspace of dimension $r = 1007$ which is tangent to X at $k \leq 63$ points, is not tangent to X elsewhere.

(ii) For $n > 5$ and $k \leq (4^n - 4^{n-3})/(3n+1)$, a general linear subspace of dimension $r = 4^n - 4^{n-3} - 1$ which is tangent to X at k points, is not tangent to X elsewhere.

(iii) For $k \leq (4^n - 4^{n-3})/(3n+1)$, then X is k-identifiable. In other words, X is k-identifiable up to $63/64 = 0.98\ldots$ of the critical (maximal) value k_c.

Proof. (i) follows from a computer-aided computation, following the the algorithm presented in Section 3.8. (ii) is a consequence of (i) and the inductive Lemma 3.2.1. (iii) follows from (ii) and Proposition 3.1.8. □

3.6 Products of three projective spaces

For the general case, in which we have projective spaces of arbitrary dimension, in order to produce examples similar to the ones of the previous section, we need a starting point for the induction.

We obtain a starting point, for the case of the product of three projective spaces $X = \mathbb{P}^a \times \mathbb{P}^b \times \mathbb{P}^c$, $2 < a \leq b \leq c$, from the following Theorem, which is due to Strassen in the case c odd (see [51], Corollary 3.7), and we generalize to any c.

The proof in [13] is apparently independent from the argument given by Strassen. Indeed, following correctly the details of the steps, one realizes that the two arguments are essentially equivalent.

Theorem 3.6.1 (V. Strassen, [51]; C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *Let X be the product of three projective spaces $X = \mathbb{P}^a \times \mathbb{P}^b \times \mathbb{P}^c$, $2 < a \leq b \leq c$, naturally embedded in $\mathbb{P}^N$, with $N = (a+1)(b+1)(c+1)-1$. Then a general linear subspace L of codimension $a+b+2$ in $\mathbb{P}^N$, that contains the span of the tangent spaces to X at k general points, with:*

$$k \leq \frac{(a+1)(b+1)(c+1)}{a+b+c+1} - c - 1,$$

is not tangent to X elsewhere.

Proof. Let $\mathbb{P}^c = \mathbb{P}(C)$, where C is a vector space of dimension $c+1$. Fix one vector $v_0 \in C$ and split C in a direct sum $C = \langle v_0 \rangle \oplus C'$, where C' is a supplementary subspace of dimension c. From the geometric point of view, this is equivalent to split the product X in two products

$$X' = \mathbb{P}^a \times \mathbb{P}^b \times \mathbb{P}^{c-1} \text{ and } X'' = \mathbb{P}^a \times \mathbb{P}^b \times \{P_0\} = \mathbb{P}^a \times \mathbb{P}^b.$$

Fix general points $P_1, \ldots, P_k \in X'$, with $P_i = v_i \otimes w_i \otimes u_i$ and let $Q_1, \ldots, Q_k$, $Q_i = v_i \otimes w_i$, be the corresponding points of X''. The linear span of the Q_i's is a space of dimension $k-1$ in $\mathbb{P}^{N''}$, where $N'' = ab + a + b$.

By assumption $k - 1 \leq N'' - \dim(X'') = N'' - a - b$. Indeed if $c + 1 \geq a + b$ then

$$k - 1 \leq \frac{(a+1)(b+1)(c+1)}{a+b+c+1} - a - b - 1 \leq (a+1)(b+1) - a - b - 1.$$

If $c + 1 < a + b$ then $k < (a+1)(b+1)/2$ and $(a+1)(b+1)/2 > a + b$.

Fix a linear space L'' of codimension $a + b + 1$ in $\mathbb{P}^{N''}$, which contains the span of the Q_i's. Since the points Q_i's are general in X'', it follows from the Theorem 2.6 in [17] (it is a generalization of the "trisecant lemma") that the linear space L'' does not meet X'' in other points. Moreover L'' is not tangent to X'' at any of the points Q_i's.

Let L' be a hyperplane in $\mathbb{P}^{N'}$, $N' = (a+1)(b+1)c - 1$, which is tangent to X' at the points P_i's. The hyperplane L' exists, since by assumption

$$k(\dim(X') + 1) < (a+1)(b+1)(c+1) - c(a+b+c) < N - 1.$$

Let L be the linear span of L' and L''. L has codimension $a + b + 2$ and it is tangent to X at the k points $P_1, \ldots, P_k$, since it contains the tangent spaces to X' at the P_i's, moreover it contains the points Q_i's, so it contains the fiber $\mathbb{P}^c$ passing through each P_i.

We want to exclude that L is tangent to X at any other point $P \neq P_i$. Call Q the projection of P to X''. If L is tangent to X at P, then it must contain the fiber $\mathbb{P}^c$ passing through P, thus it contains Q. This proves that Q is one of the Q_i's (say $Q = Q_1$), since L does not meet X'' elsewhere. But then L contains the fibers $\mathbb{P}^a$ and $\mathbb{P}^b$ at two points P, P_1 with the same projection to X''. Thus it contains these fibers at any point of

the line ℓ joining P, P_1. As ℓ contains Q_1, we get a contradiction, since $L'' = L \cap \mathbb{P}^{N''}$ is not tangent to X'' at Q_1. □

Corollary 3.6.2. *Let X be the product of three projective spaces $X = \mathbb{P}^a \times \mathbb{P}^b \times \mathbb{P}^c$, $2 < a \leq b \leq c$, naturally embedded in $\mathbb{P}^N$, with $N = (a+1)(b+1)(c+1) - 1$. Then for*

$$k \leq \frac{(a+1)(b+1)(c+1)}{a+b+c+1} - c - 1,$$

X is k-identifiable.

Proof. Follows immediately from the previous Theorem and [19]. □

The identifiability of products of three projective spaces has been studied by a long list of authors, who refined the celebrated Kruskal's bound for arbitrary tensors. We mention De Lauthawer's results for unbalanced tensor ([24]), and the general bounds found by the Chiantini and Ottaviani in [19].

It is reasonable to believe that the bound of Corollary 3.6.2, at least for some balanced case, is the best known result for tensors of type a, b, c.

3.7 Inductive bounds for the identifiability of general tensors

The same procedure we used for products of many projective lines and planes, based on the bound found in Corollary 3.6.2, can produce results for *cubic* tensors, which, in some cases, are far beyond any known result on the identifiability problem.

Then, with the above notation, we have:

Theorem 3.7.1 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *For $n \geq 3$, let X be the product of n copies of $\mathbb{P}^a$, naturally embedded in $\mathbb{P}^N$, with $N = (a+1)^n - 1$. Then for $r < (a+1)^n - (3a+1)(a+1)^{n-2}$ and for $k \leq (r+1)/(an+1)$, a general linear subspace of dimension r which is tangent to X at k points, is not tangent to X elsewhere.*

As a consequence, we get that X is k-identifiable, for

$$k \leq \frac{(a+1)^n - (3a+1)(a+1)^{n-2}}{an+1}.$$

Proof. The proof is absolutely similar to the ones of the cases $a = 1, 2, 3$ given above. We may assume $a \geq 4$. It goes by induction on $n \geq 3$, and uses Theorem 3.6.1 as a starting point.

We leave the straightforward details to the reader. □

We recall that we defined, in the introduction, the critical value

$$k_c = \frac{\prod_{i=1}^{q}(a_i+1)}{1+\sum_{i=1}^{q} a_i}$$

which is essentially the maximum for which k-identifiability can hold. Then the previous bound proves that X is k-identifiable, for

$$k \leq \frac{a(a-1)}{(a+1)^2}\, k_c.$$

Even for the case of rectangular tensors, we are able to prove *some* results, using the same procedure.

Theorem 3.7.2 (C. Bocci, L. Chiantini, G. Ottaviani, [13])**.** *Let X be the product of $q \geq 3$ projective spaces $X = \mathbb{P}^{a_1} \times \cdots \times \mathbb{P}^{a_q}$, naturally embedded in $\mathbb{P}^N$, with $N = -1 + \Pi_{i=1}^{q}(a_i+1)$. Then for*

$$r < \Pi_{i=1}^{q}(a_i+1) - (a_1+a_2+a_3+1)\Pi_{i=3}^{q}(a_i+1)$$

and for $k \leq (r+1)/(1+\sum_{i=1}^{q} a_i)$, a general linear subspace of dimension r which is tangent to X at k points, is not tangent to X elsewhere.

As a consequence, we get that X is k-identifiable, for

$$k \leq \frac{\Pi_{i=1}^{q}(a_i+1) - (a_1+a_2+a_3+1)\Pi_{i=3}^{q}(a_i+1)}{1+\sum_{i=1}^{q} a_i} = \frac{a_1a_2 - a_3}{(a_1+1)(a_2+1)}\, k_c.$$

Of course, the previous bound changes if one reorders the a_i suitably. Notice that the previous thereom requires $a_1a_2 > a_3$ in order to give a an effective range of values for k. Moreover, one of the conditions among $a_1 \gg a_3$, $a_2 \gg a_3$ and $a_1a_2 \gg a_3$ is strongly preferable to have a larger range of values for k.

We strongly believe that some *ad hoc* procedure, as well as the improvements of our computational facilities, for the starting point of the induction, are suitable to produce advancement in the inequalities of the previous results.

Let us stress that the previous bounds provide also some answers to the problem of finding the dimension of secant varieties to Segre varieties (i.e. to the dimension of paces of tensors of given rank).

Corollary 3.7.3. *Let X be the product of $q \geq 3$ projective spaces $X = \mathbb{P}^{a_1} \times \cdots \times \mathbb{P}^{a_q}$. If*

$$k \leq \frac{\Pi_{i=1}^{q}(a_i+1) - (a_1+a_2+a_3+1)\Pi_{i=3}^{q}(a_i+1)}{1+\sum_{i=1}^{q} a_i}$$

then the dimension of the k-secant variety $S_k(X)$ is the expected one, namely it is equal to $k(1+\sum_{i=1}^{q} a_i) - 1$.

We show now a list known cases when $a_q \leq \prod_{i=1}^{q-1}(a_i+1) - \left(1+\sum_{i=1}^{q-1} a_i\right)$ and the decomposition of the general tensor of rank $k < k_c$ is not unique.We refer to Section 5 of [19] for further details.

$(a_1, \ldots, a_q)$		k	number of decompositions
$(2, 3, 3)$		5	∞^1
$(2, b, b)$	b even	$\frac{3b+2}{2}$	$\infty^{\frac{b}{2}+1}$
$(1, 1, n, n)$		$2n + 1$	∞^1
$(3, 3, 3)$		6	2
$(2, 5, 5)$		8	finite, ≥ 6
$(1, 1, 1, 1, 1)$		5	2

A straightforward application of the algorithm presented in the last section shows the following

Theorem 3.7.4. *The previous list is complete for all* $(a_1, \ldots, a_q)$ *such that* $\prod_{i=1}^{q}(a_i + 1) \leq 100$.

Let us conclude this lecture by proving the non identifiability of $(\mathbb{P}^1)^5$.

Proposition 3.7.5. *The product* X *of* 5 *copies of* $\mathbb{P}^1$ *is not* 5*-identifiable. Through a general point of* $S_5(X)$ *one finds exactly two* 5*-secant,* 4*-spaces.*

Proof. Indeed, we prove that through 5 general points of X one can find an irreducible elliptic normal curve $C \subset \mathbb{P}^9$, contained in X. Since a general point of the $\mathbb{P}^9$, spanned by C, sits in exactly two subspaces of dimension 4, 5-secant to an irreducible elliptic normal curve (by [18] Proposition 5.2), it follows that the 5-th secant order of X is at least 2. In particular, X is 4-weakly defective, by [18], proposition 2.7, and the 4-th contact locus contains an elliptic normal curve as C. A computer aided computation, at 5 specific points of X, proves that indeed the 5-contact locus of X is exactly an irreducible elliptic normal curve of degree 12. The computation has been performed with the Macaulay2 Computer Algebra package [34], with the script described in [12]. Thus 5-th secant order of X is 2 (by Theorem 3.1.5) and the claim is proved.

To prove the existence of the curve C passing through 5 general points $P_1, \ldots, P_5$ of X, we start with the product of three lines $X' = \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$. Through the 5 points $P'_1, \ldots, P'_5 \in X'$, projection of the P_i's, one can find a 2-dimensional family $\mathcal{F}$ of elliptic normal curves C' of degree 6. Indeed $X' \subset \mathbb{P}^5$ is a sestic threefold with elliptic curve sections, and there is a 2-dimensional family of hyperplanes passing through 5

general points of X. $\mathcal{F}$ is parametrized by points of some plane Π, obtained by projecting $\mathbb{P}^5$ from the span of the P'_i's.

Consider now the product X'' of the two remaining copies of $\mathbb{P}^1$, so that $X = X' \times X''$. We also get 5 distinguished general points $P''_1, \ldots, P''_5 \in X''$. For any curve C' of the family $\mathcal{F}$, we have a 7-dimensional family of embeddings $C' \to X''$. Thus, adding the automorphisms of C', for $C' \in \mathcal{F}$ general, we may assume that each P'_i, $i = 1, \ldots, 4$, goes to the corresponding P''_i. The condition that P'_5 goes to P''_5 determines two algebraic condition on the family, hence two algebraic curves on Π. Thus, there is at least one curve C' of the family, for which P'_5 goes to P''_5. This determines an elliptic normal curve C in X, passing through the 5 given general points P_i's. The fact that C is irreducible, for a general choice of the points, follows by the computer-aided computation, on a specific example. □

3.8 The algorithm

The algorithm we have used has been implemented in Macaulay2 [34] and it can be found as ancillary file in the arXiv submission of the paper [13].

The steps are the following.

1. We choose s random points $p_1, \ldots, p_s$ on the Segre variety X, working on an affine chart. The point p_1 can be chosen as $(1, 0, \ldots)$ on each factor.

2. We compute the equations of the span of tangent spaces $< T_{p_1}, \ldots, T_{p_s} >$.

3. For any of the cartesian equations we compute its partial derivatives, the common locus is the locus C of points p such that $T_p X \subset \langle T_{p_1}, \ldots, T_{p_s} \rangle$.

4. We compute the rank of the jacobian matrix of C at p_1. If it is equal to the dimension of X then X is k-identifiable. If it is smaller than the dimension of X then a further analysis is required.

Lecture 4. Quasi-independence models on matrices and tensors

Let us study now some class of statistical models on contingency tables and its generalization on tensors. The aim of the lecture is to show combinatorial approaches to the research of model invariants.

4.1 Diagonal-effect models

Diagonal-effect models for square $I \times I$ tables can be defined in at least two ways. In the field of toric models, one can define these models in monomial form as follows.

Definition 4.1.1. *The diagonal-effect model $\mathcal{M}_1$ is defined as the set of probability matrices $P \in \Delta$ such that:*

$$p_{ij} = r_i c_j \ \text{for} \ i \neq j \tag{4.32}$$

and

$$p_{ij} = r_i c_j \gamma_i \ \text{for} \ i = j \tag{4.33}$$

where r, c and γ are non-negative vectors with length I.

In literature, such a model is also known as quasi-independence model, see [2]. As the model in Definition 4.1.1 is a toric model, it is relatively easy to find the invariants. Eliminating the parameters r, c and γ one obtains the following result.

Theorem 4.1.2. *The invariants of the model $\mathcal{M}_1$ are the binomials*

$$p_{ij}p_{i'j'} - p_{ij'}p_{i'j} \tag{4.34}$$

for i, i', j, j' all distinct, and

$$p_{ii'}p_{i'i''}p_{i''i} - p_{ii''}p_{i''i'}p_{i'i} \tag{4.35}$$

for i, i', i'' all distinct.

In the framework of the mixture models, the diagonal-effect models have an alternative definition as follows.

Definition 4.1.3. *The diagonal-effect model $\mathcal{M}_2$ is defined as the set of probability matrices P such that*

$$P = \alpha c r^t + (1-\alpha)D \tag{4.36}$$

where r and c are non-negative vectors with length I and sum 1, $D = \operatorname{diag}(d_1, \dots, d_I)$ is a non-negative diagonal matrix with sum 1, and $\alpha \in [0, 1]$.

Remark 4.1.4. Notice that while in Definition 4.1.1 the normalization is applied once, in Definition 4.1.3 the normalization is applied twice as we require that both cr^t and D are probability matrices. This difference will be particularly relevant in the study of the geometry of the models.

We study the invariants and some geometrical properties of these models.

Theorem 4.1.5 (C. Bocci, E. Carlini, F. Rapallo, [10])**.** *The models $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same invariants.*

Proof. Writing explicitly the polynomials in Equations (4.32) and (4.33) it is easy to check that each γ_i appears in only one polynomial. The same for each d_i in Equations (4.36). Thus, following Theorem 3.4.5 in [41], such polynomials are deleted when we eliminate the indeterminates γ_i's and d_i's.

As the remaining polynomials, corresponding to off-diagonal cells, are the same in both models, the models $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same invariants. □

In order to study in more details the connections between $\mathcal{M}_1$ and $\mathcal{M}_2$ we further investigate their geometric structure. The non-negativity conditions imposed in the definitions imply that $\mathcal{M}_1 \neq \mathcal{M}_2$ and neither $\mathcal{M}_2 \subset \mathcal{M}_1$ nor $\mathcal{M}_1 \subset \mathcal{M}_2$. We can show this by two easy examples.

First, let r and c respectively the vectors, of length I, $(\frac{1}{I}, \frac{1}{I}, \dots, \frac{1}{I})$ and $(\frac{1}{I-1}, \frac{1}{I-1}, \dots, \frac{1}{I-1})$ and define γ as the zero vector. Thus, the probability table we obtain in toric form is:

$$P = \begin{pmatrix} 0 & \frac{1}{I(I-1)} & \frac{1}{I(I-1)} & \cdots & \frac{1}{I(I-1)} \\ \frac{1}{I(I-1)} & 0 & \frac{1}{I(I-1)} & \cdots & \frac{1}{I(I-1)} \\ \frac{1}{I(I-1)} & \frac{1}{I(I-1)} & 0 & \cdots & \frac{1}{I(I-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{I(I-1)} & \frac{1}{I(I-1)} & \frac{1}{I(I-1)} & \cdots & 0 \end{pmatrix}.$$

Such probability matrix belongs to $\mathcal{M}_1$ by constructions, while it does not belong to $\mathcal{M}_2$. In fact, $p_{11} = 0$ in Equation (4.36) would imply either $\alpha = 0$ (a contradiction, as P is not a diagonal matrix), or $r_1 = 0$ (a

contradiction, as P has not the first row with all 0's), or $c_1 = 0$ (a contradiction, as P has not the first column with all 0's).

On the other hand, let P be the diagonal matrix

$$P = \begin{pmatrix} \frac{1}{I} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{I} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{I} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{I} \end{pmatrix}.$$

Such probability matrix belongs to $\mathcal{M}_2$ by setting $\alpha = 0$ and $D = \text{diag}(\frac{1}{I}, \dots, \frac{1}{I})$, while it does not belong to $\mathcal{M}_1$. To prove this it is enough to note that $p_{12} = 0$ would imply either $r_1 = 0$ (a contradiction, as the first row of P is not zero), or $c_2 = 0$ (a contradiction, as the second column of P is not zero).

Nevertheless, in the open simplex we can prove one of the inclusions.

Proposition 4.1.6 (C. Bocci, E. Carlini, F. Rapallo, [10])**.** *In the open simplex* $\Delta_{>0}$*,*

$$\mathcal{M}_2 \subset \mathcal{M}_1 \tag{4.37}$$

Proof. In fact, let us consider a probability table in $\mathcal{M}_2$, given by $P = \alpha c r^t + (1-\alpha)D$. As $P \in \Delta_{>0}$, $\alpha \neq 0$, $r_i \neq 0$ for all $i = 1, \dots, I$ and $c_j \neq 0$ for all $j = 1, \dots, I$. Then we can describe P as an element of $\mathcal{M}_1$ in the following way. We define $r_i = r_i$ for all i and $c_j = \alpha c_j$, for all j. After that, it is enough to find the diagonal parameters by solving the equations

$$\alpha r_i c_i \gamma_i = \alpha r_i c_i + (1-\alpha) d_i$$

that is, as $\alpha \neq 0$, $r_i \neq 0$, and $c_i \neq 0$, we have

$$\gamma_i = 1 + \frac{(1-\alpha)d_i}{\alpha r_i c_i}.$$

□

Moreover, in the open simplex $\Delta_{>0}$, the inclusion in Proposition 4.1.6 is strict. Let us analyze the probability matrices in the difference $\mathcal{M}_1 \setminus \mathcal{M}_2$.

Consider three vectors $r = (r_1, \dots, r_I)$, $c = (c_1, \dots, c_I)$ and $\gamma = (\gamma_1, \dots, \gamma_I)$. Using these vectors, we define the probability table P as in Definition 4.1.1 and then we normalize it, i.e. dividing by $N_T = \sum_{i \neq j} r_i c_j + \sum_{i=j} r_i c_j \gamma_i$. Define also $N = \sum_{ij} r_i c_j$ (which can be seen as the normalization of the toric model when γ is the unit vector, i.e., it is the vector with all components equal to one).

We want to find three vectors $\overline{c} = (\overline{c}_1, \dots, \overline{c}_I)$, $\overline{r} = (\overline{r}_1, \dots, \overline{r}_I)$, $d = (d_1, \dots, d_I)$, with $\sum \overline{r}_i = \sum \overline{c}_i = \sum d_i = 1$ and a scalar $0 \leq \alpha \leq 1$ such that

$$\begin{aligned} \frac{1}{N_T} \begin{pmatrix} r_1c_1\gamma_1 & r_1c_2 & r_1c_3 & \dots & r_1c_I \\ r_2c_1 & r_2c_2\gamma_2 & r_2c_3 & \dots & r_2c_I \\ r_3c_1 & r_3c_2 & r_3c_3\gamma_3 & \dots & r_3c_I \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_Ic_1 & r_Ic_2 & r_Ic_3 & \dots & r_Ic_I\gamma_I \end{pmatrix} = \\ = \alpha \begin{pmatrix} \overline{r}_1\overline{c}_1 & \overline{r}_1\overline{c}_2 & \overline{r}_1\overline{c}_3 & \dots & \overline{r}_1\overline{c}_I \\ \overline{r}_2\overline{c}_1 & \overline{r}_2\overline{c}_2 & \overline{r}_2\overline{c}_3 & \dots & \overline{r}_2\overline{c}_I \\ \overline{r}_3\overline{c}_1 & \overline{r}_3\overline{c}_2 & \overline{r}_3\overline{c}_3 & \dots & \overline{r}_3\overline{c}_I \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{r}_I\overline{c}_1 & \overline{r}_I\overline{c}_2 & \overline{r}_I\overline{c}_3 & \dots & \overline{r}_I\overline{c}_I \end{pmatrix} + (1-\alpha) \begin{pmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & 0 & d_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & d_I \end{pmatrix}. \end{aligned} \tag{4.38}$$

We start studying the off-diagonal elements. Consider first the case $N_T > N$. Thus we have $\frac{r_ic_j}{N_T} < \frac{r_ic_j}{N}$ and $\frac{N}{N_T} < 1$. In this situation the only possible choice is given by

$$\alpha = \frac{N}{N_T} \quad \overline{r}_i = \frac{r_i}{\sum r_i}, \quad \overline{c}_j = \frac{c_j}{\sum c_j}. \tag{4.39}$$

In fact, recalling the definition of N, we have

$$\alpha \overline{r}_i \overline{c}_j = \frac{N}{N_T} \frac{r_i}{\sum r_i} \frac{c_j}{\sum c_j} = \frac{N}{N_T} \frac{r_ic_j}{N} = \frac{r_ic_j}{N_T} \tag{4.40}$$

for all i, j with $i \neq j$. Taking the log-probabilities, we obtain a linear system. It is easy to prove, as in Chapter 6 of [47], that the rank of this system is equal to $(2I - 1)$. Hence, considering the normalizing equations for $\overline{r}$ and $\overline{c}$, we see that the solution in (4.39) is unique.

Let us consider the generic equation of the $i-$th diagonal element:

$$r_ic_i\gamma_i = \alpha \overline{r}_i\overline{c}_i + (1-\alpha)d_i\,.$$

After substituting the previous values for $\overline{r}_i$, $\overline{c}_i$ and α we get

$$r_ic_i\gamma_i = \frac{N}{N_T} \frac{r_ic_i}{N} + \frac{N_T - N}{N_T} d_i\,.$$

As we consider matrices in $\Delta_{>0}$, the quantity r_ic_i is different from zero. Therefore, after multiplying for N_T and dividing by r_ic_i we obtain

$$\gamma_i = 1 + \frac{N_T - N}{r_ic_i} d_i$$

that is

$$d_i = \frac{r_i c_i}{N_T - N}(\gamma_i - 1)$$

Thus we see that the $P \in \mathcal{M}_1 \setminus \mathcal{M}_2$ when $N_T > N$ and there exists at least an index i such that $\gamma_i < 1$.

When $N_T = N$, from Equations (4.40) we obtain $\alpha = 1$. Therefore in Equation (4.38) the matrix on the right hand side has rank 1, and this implies that $P \in \mathcal{M}_2$ if and only if $\gamma_i = 1$ for all i.

Consider now the case $N_T < N$. Hence we have $\frac{r_i c_j}{N_T} > \frac{r_i c_j}{N}$ and $\frac{N}{N_T} > 1$. Again the only possible choice for the off-diagonal elements would be given by

$$\alpha = \frac{N}{N_T}, \quad \bar{r}_i = \frac{r_i}{\sum r_i}, \quad \bar{c}_i = \frac{c_i}{\sum c_i},$$

but in this case $\alpha = \frac{N}{N_T} > 1$. Thus we conclude that all $P \in \mathcal{M}_1$ with $N_T < N$ are in $\mathcal{M}_1 \setminus \mathcal{M}_2$. This leads to the following result.

Theorem 4.1.7 (C. Bocci, E. Carlini, F. Rapallo, [10])**.** *Let $P \in \mathcal{M}_1 \cap \Delta_{>0}$ be a strictly positive probability table given by the vectors $r = (r_1, \dots, r_I)$, $c = (c_1, \dots, c_I)$ and $\gamma = (\gamma_1, \dots, \gamma_I)$. Define $N_T = \sum_{i \neq j} r_i c_j + \sum_{i=j} r_i c_j \gamma_i$ and $N = \sum_{ij} r_i c_j$. Then $P \in \mathcal{M}_1 \setminus \mathcal{M}_2$ if one of the following situations holds:*

(i) $N_T < N$;

(ii) $N_T = N$ *and there exists at least an index* i *such that* $\gamma_i \neq 1$;

(iii) $N_T > N$ *and there exists at least an index* i *such that* $\gamma_i < 1$.

4.2 A geometric description of the diagonal-effect models

In this section, we try to describe the models we studied using some geometric flavor. This analysis will also shed some light on the elements in $\mathcal{M}_1 \setminus \mathcal{M}_2$. We use very basic and classic geometric ideas and facts. As references, we suggest [35] and [36].

We start with the model $\mathcal{M}_1$. The basic object we need is the variety V describing all $I \times I$ matrices having rank at most one. When we fix $\gamma_i = 1, i = 1, \dots, I$ the parametrization in (4.32) and (4.33) is just describing V. Hence, fixing values for all the c_i's and the r_i's and setting $\gamma_j = 1, j = 1, \dots, I$ we obtain a point $M \in V$. Now, if we let γ_l to vary we are describing a line passing through M and moving in the direction of the vector $(0, \dots, 1, \dots, 0)$, where the only non zero coordinate is the (l, l)-th; the set of all these lines is a cylinder. Now we set $\gamma_l = a\zeta$ and $\gamma_m = b\zeta$ for fixed reals a and b. When we let ζ vary, we are now describing a cylinder with directrix parallel to the line of equations

$$\begin{cases} b p_{ll} - a p_{mm} = 0 \\ p_{ij} = 0 \end{cases}$$

for $(i, j) \neq (l, l), (m, m)$. The same argument can be repeated fixing linear relations among the diagonal elements. In conclusion, we can describe $\mathcal{M}_1$ as the intersection of the simplex with the union of cylinders having base V and directrix parallel to the directions given by diagonal elements.

We now use the join of two varieties, i.e. the closure of the set of all the lines joining a point of any variety with any point of another variety. In order to do this, we also need to consider W the variety of diagonal matrices. Then $\mathcal{M}_2$ is the union of the segment joining a point of $V \cap \Delta$ with a point of $W \cap \Delta$, i.e. a subvariety of the join of V and W. Each of this segment lies on a line contained in one of the cylinder we used to construct $\mathcal{M}_1$. Hence we get again the inclusion $\mathcal{M}_2 \subset \mathcal{M}_1$ in Δ.

4.3 Common-diagonal-effect models

A different version of the diagonal-effect models are the so-called common-diagonal-effect models. The definitions are as in the models above but:

- The entries of the vector γ are all equal in the toric model definition;
- The matrix D is $\mathrm{diag}(\frac{1}{I}, \ldots, \frac{1}{I})$ in the mixture model definition.

This kind of models is much more complicated than the models in Section 4.1. Just to have a first look at these models, we note that for $I = 3$ the diagonal-effect models have only one invariant. For the common-diagonal-effect models, we have computed the invariants with CoCoA, see [20], for $I = 3$ and we have obtained the following lists of invariants.

For the toric model we obtain 9 binomials:

$$p_{12}p_{23}p_{31} - p_{13}p_{21}p_{32}\,,$$

$$p_{13}p_{22}p_{31} - p_{11}p_{23}p_{32}\,,$$

$$-p_{11}p_{23}p_{32} + p_{12}p_{21}p_{33}\,,$$

$$-p_{22}p_{23}p_{31}^2 + p_{21}^2p_{32}p_{33}\,,$$

$$p_{12}p_{22}p_{31}^2 - p_{11}p_{21}p_{32}^2\,,$$

$$-p_{11}p_{13}p_{32}^2 + p_{12}^2p_{31}p_{33}\,,$$

$$-p_{13}^2p_{22}p_{32} + p_{12}^2p_{23}p_{33}\,,$$

$$-p_{11}p_{23}^2p_{31} + p_{13}p_{21}^2p_{33}\,,$$

$$p_{13}^2p_{21}p_{22} - p_{11}p_{12}p_{23}^2\,.$$

For the mixture model we obtain:

- 1 binomial

$$p_{12}p_{23}p_{31} - p_{13}p_{21}p_{32}\,;$$

- 12 polynomials with 4 terms

$$p_{13}p_{21}p_{22} - p_{12}p_{21}p_{23} + p_{13}p_{23}p_{31} - p_{13}p_{21}p_{33}\,,$$

$$-p_{12}p_{13}p_{22} + p_{12}^2p_{23} - p_{13}^2p_{32} + p_{12}p_{13}p_{33}\,,$$

$$p_{13}p_{21}p_{31} - p_{11}p_{23}p_{31} + p_{22}p_{23}p_{31} - p_{21}p_{23}p_{32}\,,$$

$$p_{12}p_{13}p_{31} - p_{11}p_{13}p_{32} + p_{13}p_{22}p_{32} - p_{12}p_{23}p_{32}\,,$$

$$p_{13}p_{21}^2 - p_{11}p_{21}p_{23} - p_{23}^2p_{31} + p_{21}p_{23}p_{33}\,,$$

$$p_{13}^2p_{21} - p_{11}p_{13}p_{23} + p_{13}p_{22}p_{23} - p_{12}p_{23}^2\,,$$

$$p_{12}p_{13}p_{21} - p_{11}p_{12}p_{23} - p_{13}p_{23}p_{32} + p_{12}p_{23}p_{33}\,,$$

$$-p_{21}p_{22}p_{31} - p_{23}p_{31}^2 + p_{21}^2p_{32} + p_{21}p_{31}p_{33}\,,$$

$$-p_{12}p_{22}p_{31} + p_{12}p_{21}p_{32} - p_{13}p_{31}p_{32} + p_{12}p_{31}p_{33}\,,$$

$$p_{12}p_{31}^2 - p_{11}p_{31}p_{32} - p_{22}p_{31}p_{32} - p_{21}p_{32}^2\,,$$

$$p_{12}p_{21}p_{31} - p_{11}p_{21}p_{32} - p_{23}p_{31}p_{32} + p_{21}p_{32}p_{33}\,,$$

$$p_{12}^2p_{31} - p_{11}p_{12}p_{32} - p_{13}p_{32}^2 + p_{12}p_{32}p_{33}\,;$$

- 6 polynomials with 8 terms

$$p_{11}p_{13}p_{22} - p_{13}p_{22}^2 - p_{11}p_{12}p_{23} + p_{12}p_{22}p_{23} + \\ + p_{13}^2p_{31} - p_{13}p_{23}p_{32} - p_{11}p_{13}p_{33} + p_{13}p_{22}p_{33}\,,$$

$$p_{11}p_{13}p_{21} - p_{11}^2p_{23} - p_{12}p_{21}p_{23} + p_{11}p_{22}p_{23} + \\ + p_{23}^2p_{32} - p_{13}p_{21}p_{33} + p_{11}p_{23}p_{33} - p_{22}p_{23}p_{33}\,,$$

$$-p_{11}p_{22}p_{31} + p_{22}^2p_{31} - p_{13}p_{31}^2 + p_{11}p_{21}p_{32} + \\ - p_{21}p_{22}p_{32} + p_{23}p_{31}p_{32} + p_{11}p_{31}p_{33} - p_{22}p_{31}p_{33}\,,$$

$$p_{11}p_{12}p_{31} - p_{11}^2p_{32} - p_{12}p_{21}p_{32} + p_{11}p_{22}p_{32} + \\ + p_{23}p_{32}^2 - p_{12}p_{31}p_{33} + p_{11}p_{32}p_{33} - p_{22}p_{32}p_{33}\,,$$

$$p_{12}p_{21}^2 - p_{11}p_{21}p_{22} - p_{11}p_{23}p_{31} - p_{21}p_{23}p_{32} + \\ + p_{11}p_{21}p_{33} + p_{21}p_{22}p_{33} + p_{23}p_{31}p_{33} - p_{21}p_{33}^2\,,$$

$$p_{12}^2p_{21} - p_{11}p_{12}p_{22} - p_{11}p_{13}p_{32} - p_{12}p_{23}p_{32} + \\ + p_{11}p_{12}p_{33} + p_{12}p_{22}p_{33} + p_{13}p_{32}p_{33} - p_{12}p_{33}^2\,;$$

- 1 polynomial with 12 terms

$$p_{11}p_{12}p_{21} - p_{11}^2p_{22} - p_{12}p_{21}p_{22} + p_{11}p_{22}^2 + \\ - p_{11}p_{13}p_{31} + p_{22}p_{23}p_{32} + p_{11}^2p_{33} - p_{22}^2p_{33} + \\ + p_{13}p_{31}p_{33} - p_{23}p_{32}p_{33} - p_{11}p_{33}^2 + p_{22}p_{33}^2\,.$$

Therefore, as in Theorem 4.1.2, we can easily derive the invariants. We do not write explicitly the analog of Theorem 4.1.2 for common-diagonal-effect models in order to save space.

The study of the common-diagonal-effect models in mixture form is much more complicated. In fact, notice that in the computations above, the mixture model present invariants which are not binomials. However, some partial results can be stated.

Theorem 4.3.1 (C. Bocci, E. Carlini, F. Rapallo, [10]). *Define the following polynomials:*

(a) For i, j, k, l all distinct we define

$$b_{ijkl} = p_{ij}p_{kl} - p_{il}p_{kj}\,;$$

(b) For i, j, k, all distinct we define

$$t_{ijk} = p_{ij}p_{jk}p_{ki} - p_{ik}p_{kj}p_{ji}\,;$$

(c) For (i,j) and (k,l) two distinct pairs in $\{1,\dots,I\}$ with $i \neq j$, and $k \neq l$ and $m \in \{1,\dots,I\} \setminus \{i,j\}$ and $n \in \{1,\dots,I\} \setminus \{k,l\}$ with $m \neq n$ we define

$$f_{ijklmn} = p_{ij}p_{kl}p_{nn} - p_{ij}p_{nl}p_{kn} - p_{ij}p_{kl}p_{mm} + p_{kl}p_{mj}p_{im}\,;$$

(d) for two distinct indices i and j in $\{1,\dots,I\}$ and for $k \in \{1,\dots,I\} \setminus \{i,j\}$ we define

$$g_{ijk} = p_{ij}p_{ii}p_{kk} + p_{ij}p_{jj}p_{kk} - p_{ij}p_{ii}p_{jj} + p_{ij}p_{kk}^2 + \\ + p_{kk}p_{ik}p_{kj} - p_{ii}p_{ik}p_{kj} + p_{ij}^2p_{ji} - p_{ij}p_{kj}p_{jk}\,;$$

(e) For i, j, k, all distinct we define

$$h_{ijk} = p_{ii}p_{jj}^2 + p_{ii}^2 p_{kk} + p_{jj}p_{kk}^2 - p_{ii}^2 p_{jj} - p_{jj}^2 p_{kk} - p_{ii}p_{kk}^2 + p_{ii}p_{ij}p_{ji} + \\ - p_{ii}p_{ik}p_{ki} + p_{jj}p_{jk}p_{kj} - p_{jj}p_{ji}p_{ij} + p_{kk}p_{ki}p_{ik} - p_{kk}p_{kj}p_{jk}\,.$$

Then the previous polynomials are invariants for the common-diagonal-effect models in mixture form.

Proof. Cases (a) and (b) follow from Theorem 4.1.2 since the off-diagonal elements of the probability table are described, up to scalar, in the same monomial form as for the elements of $\mathcal{M}_1$.

For case (c), consider the term $g_1 = p_{ij}p_{kl}p_{nn}$ in f_{ijklmn}. This gives two monomials: $\alpha^3 r_i c_j r_k c_l r_n c_n$ and $\alpha^2 r_i c_j r_k c_l (1-\alpha)d$, where $d = 1/I$. The term $-g_2 = -p_{ij}p_{nl}p_{kn}$ of f_{ijklmn} cancels the first monomial of g_1. In fact $-p_{ij}p_{nl}p_{kn} = \alpha^3 r_i c_j r_n c_l r_k c_n$. Since in g_2 there are not diagonal variables, we need another term in order to cancel the second monomial of g_1. Thus we subtract, to $g_1 - g_2$, a term of the form $g_3 = p_{ij}p_{kl}p_{mm}$ which gives the monomials $-\alpha^2 r_i c_j r_k c_l (1-\alpha)d$ and $-\alpha^3 r_i c_j r_k c_l r_m c_m$. To cancel this last monomial it is enough to add the term $g_4 = p_{kl}p_{mj}p_{im} = \alpha^3 r_k c_l r_m c_j r_i c_m$. Thus $f_{ijklmn} = g_1 - g_2 - g_3 + g_4$ vanishes on the entries of a probability table of the mixture model with common diagonal effect.

For case (d), consider first the terms with pairs of variables on the diagonal.

$$p_{ij}p_{ii}p_{kk} = \alpha^3 r_i^2 r_k c_i c_j c_k + \alpha^2 r_i^2 c_i c_j d - \alpha^3 r_i^2 c_i c_j d + \boxed{\alpha^2 r_i r_k c_j c_k d} + \\ + \alpha r_i c_j d^2 - 2\alpha^2 r_i c_j d^2 - \alpha^3 r_i r_k c_j c_k d + \alpha^3 r_i c_j d^2\,;$$

$$p_{ij}p_{jj}p_{kk} = \alpha^3 r_i r_j r_k c_j^2 c_k + \alpha^2 r_i r_j c_j^2 d - \alpha^3 r_i r_j c_j^2 d + \alpha^2 r_i r_k c_j c_k d + \\ + \alpha r_i c_j d^2 - 2\alpha^2 r_i c_j d^2 - \boxed{\alpha^3 r_i r_k c_j c_k d} + \alpha^3 r_i c_j d^2\,;$$

$$p_{ij}p_{ii}p_{jj} = \alpha^3 r_i^2 r_j c_i c_j^2 + \alpha^2 r_i^2 c_i c_j d - \alpha^3 r_i^2 c_i c_j d + \alpha^2 r_i r_j c_j^2 d + \\ + \alpha r_i c_j d^2 - 2\alpha^2 r_i c_j d^2 - \alpha^3 r_i r_j c_j^2 d + \alpha^3 r_i c_j d^2\,;$$

$$p_{ij}p_{kk}^2 = \alpha^3 r_i r_k^2 c_j c_k^2 + 2\alpha^2 r_i r_k c_j c_k d - \boxed{2\alpha^3 r_i r_k c_j c_k d} + \alpha r_i c_j d^2 + \\ - 2\alpha^2 r_i c_j d^2 + \alpha^3 r_i c_j d^2\,.$$

It is easy to see that while some terms, such as $\alpha^3 r_i c_j d^2$, are simply cancelled considering the difference of two monomials, other terms, such as the boxed ones, appear in different monomials. However, they appear with the appropriate coefficients and considering $p_{ij}p_{ii}p_{kk} + p_{ij}p_{jj}p_{kk} - p_{ij}p_{ii}p_{jj} - p_{ij}p_{kk}^2$ we cancel most of them. In fact we obtain

$$\alpha^3 r_i^2 r_k c_i c_j c_k - \alpha^3 r_i r_k^2 c_j c_k^2 - \alpha^3 r_i^2 r_j c_j^2 c_i + \alpha^3 r_i r_j r_k c_j^2 c_k\,.$$

The only possible way to cancel the term $-\alpha^3 r_i r_k^2 c_j c_k^2$ is to add the monomial $p_{ik}p_{kj}p_{kk} = \alpha^3 r_i r_k^2 c_j c_k^2 + \alpha^2 r_i r_k c_j c_k d - \alpha^3 r_i r_k c_j c_k d$. However this monomial adds two more terms that can be cancelled by using another monomial with a variable in the diagonal, that is $p_{ii}p_{ik}p_{kj} = \alpha^3 r_i^2 r_k c_j c_j c_k + \alpha^2 r_i r_k c_j c_k d - \alpha^3 r_i r_k c_j c_k d$. After that, the only two missing terms are $-\alpha^3 r_i^2 r_j c_j^2 c_i + \alpha^3 r_i r_j r_k c_j^2 c_k$ which can be cancelled by adding $p_{ij}^2 p_{ji} - p_{ij}p_{kj}p_{jk}$.

For the case (e), we omit the complete details of the proof. One has to proceed as in cases (c) and (d) considering separately $p_{ii}p_{jj}^2 + p_{ii}^2 p_{kk} + p_{jj}p_{kk}^2 - p_{ii}^2 p_{jj} - p_{jj}^2 p_{kk} - p_{ii}p_{kk}^2$ and the contributions of $p_{ii}p_{ij}p_{ji} - p_{ii}p_{ik}p_{ki}$, $p_{jj}p_{jk}p_{kj} - p_{jj}p_{ji}p_{ij}$ and $p_{kk}p_{ki}p_{ik} - p_{kk}p_{kj}p_{jk}$. □

With some computations with CoCoA, we have found that the polynomials defined in Theorem 4.3.1 define the model $\mathcal{M}_2$ for $I = 3, 4, 5$. We conjecture that this fact is true in general.

4.4 Quasi-independence models on tensors

The quasi-independence models studied in the previous sections can be introduced for a bigger number of random variables. In this case we substitute the probability matrix with a probability tensors. We make the assumption that the cardinality I of state space for each variable is the same. We use now the lower-scripts n and I to denote that the model has n variables with state space $[I]$. The Definitions 4.1.1 and 4.1.3 change in the following way

Definition 4.4.1. *The diagonal-effect model $\mathcal{M}_{1,n,I}$ is defined as the set of probability tensors $P \in \Delta$ such that:*

$$p_{i_1 i_2 \cdots i_n} = v_{1i_1} v_{2i_2} \cdots v_{ni_n} \text{ for at least two disting indices among } i_1, i_2, \ldots i_n \tag{4.41}$$

and

$$p_{i_1 i_2 \cdots i_n} = v_{1i_1} v_{2i_2} \cdots v_{ni_n} \gamma_{i_1} \text{ for } i_1 = i_2 = \cdots = i_n \tag{4.42}$$

where $v_1, v_2, \ldots, v_n$ and γ are non-negative vectors with length I.

Definition 4.4.2. *The diagonal-effect model $\mathcal{M}_{2,n,I}$ is defined as the set of probability matrices P such that*

$$P = \alpha v_1 \otimes v_2 \otimes \cdots \otimes v_n + (1 - \alpha) D \tag{4.43}$$

where $v_1, v_2, \ldots, v_n$ are non-negative vectors with length I and sum 1, $D = \operatorname{diag}(d_1, \ldots, d_I)$ is a non-negative diagonal tensor with sum 1, and $\alpha \in [0, 1]$.

Again we have the following

Theorem 4.4.3 (C. Bocci, M. Benedettelli, F. Rapallo, [8])**.** *The models $M_{1,n,I}$ and $M_{2,n,I}$ have the same invariants.*

Proof. left to the reader. □

Example 4.4.4. Let $I = 2$ and $n = 3$, then the invariants of $M_{1,3,2}$ are precisely:

$$P_{121}P_{212} - P_{112}P_{221}$$

$$P_{122}P_{211} - P_{121}P_{212}$$

From Theorem 4.1.5 we know that they are also the invariants for $M_{2,3,2}$.

For $n > 2$ we get new invariants which are binomials of degree d built in the following way: we choose many non-diagonal terms (repetition allowed) as the chosen degree d and we perform $d - 1$ changes in the indices in such a way diagonal elements never appear.

Example 4.4.5. For $n = 3$, and $d = 2$ we have the following invariant:

$$P_{122}P_{131} - P_{121}P_{132}$$

The monomial have a change in the last index.

For $d = 3$ we have the following invariant:

$$P_{211}P_{233}P_{313} - P_{213}^2P_{331}.$$

Here we have two consecutive changes: between P_{211}, P_{313} and between P_{233}, P_{311} (taking in mind that a change is already done).

4.5 Common-diagonal-effect models on tensors

We fix our attention in the common-diagonal-effect models. Here we show a way to built, combinatorically, new invariants, without using Elimination Theory. The definitions for the common-diagonal-effect models on tensors are the same of the case for matrices:

- The entries of the vector γ are all equal in the toric model definition. We denote this model by $\widetilde{M}_{1,n,I}$.
- The tensor D is $\mathrm{diag}(\frac{1}{I}, \ldots, \frac{1}{I})$ in the mixture model definition. We denote this model by $\widetilde{M}_{2,n,I}$.

If we try to compute the invariants for these two models, we see that the list is quite big. However, as already said, it is possible to compute them using a combinatorial approach. We start with some definitions and remarks.

Consider an $m \times n-$matrix A:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

where the columns denote the number of the indices (recall that for each probability $p_{i_1 i_2 \cdots i_n}$ we have n indices $i_1, i_2, \dots, i_n$) and the number of rows can vary according to the degree of the chosen monomial. For example, a_{11} denote the value of the first index (i_1 in our case) in the first element of the monomial and so on.

Denote with $\rho(A)$ the number of rows with identical entries, i.e.

$$\rho(A) = \sharp\{t : a_{t1} = a_{t2} = \cdots = a_{tn}\}.$$

Thus $\rho(A)$ represents the number of diagonal elements in the monomial.

Example 4.5.1. Consider the monomial $p_{123}p_{121}p_{333}$ (then $n = 3$). The matrix associated to the monomial is

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 1 \\ 3 & 3 & 3 \end{pmatrix}.$$

If we consider, instead, the monomial, $p_{121}^3 p_{333}$ (then $n = 3$) then its associated matrix

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \\ 3 & 3 & 3 \end{pmatrix}$$

that is, we consider three occurrence of (1 2 1) according to the degree of p_{121}. In both example $\rho(A) = 1$.

Consider now an element $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n) \in S_m^n \setminus \Delta$, where S_m is the group of permutation on m elements. σ is composed by n permutation $\sigma_i \in S_m$, $i = 1, 2, \dots, n$, such that a least two of them are distinct. We can associate a map to σ

$$F_\sigma : M_{m\times n} \to M_{m\times n}$$

which sends A in

$$F_\sigma(A) = \begin{pmatrix} a_{\sigma_1(1)1} & a_{\sigma_2(1)2} & \dots & a_{\sigma_n(1)n} \\ a_{\sigma_1(2)1} & a_{\sigma_2(2)2} & \dots & a_{\sigma_n(2)n} \\ \vdots & \vdots & & \vdots \\ a_{\sigma_1(m)1} & a_{\sigma_2(m)2} & \dots & a_{\sigma_n(m)n} \end{pmatrix}.$$

Let R_{ts} be the $m \times m-$matrix with enries $r_{ts} = r_{st} = 1$ and 0 otherwise. If we multiply (at right) a $m \times p-$matrix M for R_{ts} we get a new $m \times p-$matrix equal to M but with $t-$th and $s-$th rows permuted. We denote by $R_{t_1 t_2 \dots t_r}$ the composition of matrices of the form R_{ts}

Definition 4.5.2. *Let A be a $m \times n-$matrix. We define the two $\sigma-$***sets** *of A as*

$$\Sigma(A) = \{\sigma(A) : \sigma \in S_m^n \setminus \Delta \ : \ \rho(\sigma(A)) = \rho(A) \text{ and } \nexists R_{t_1 t_2 \dots t_r} \text{ with } R_{t_1 t_2 \dots t_r}\sigma(A) = A\}.$$

$$\Sigma'(A) = \{\sigma(A) : \sigma \in S_m^n \setminus \Delta \ : \ \rho(\sigma(A)) \leq \rho(A) \text{ and } \nexists R_{t_1 t_2 \dots t_r} \text{ with } R_{t_1 t_2 \dots t_r}\sigma(A) = A\}.$$

The idea is clear:

- we consider a monomial $P = \Pi_{\{i_1, i_2, \dots i_n\} \subset I^n} p_{i_1 i_2 \dots i_n}^{\alpha_{i_1 i_2 \dots i_n}}$
- we associate to P its $t \times n-$matrix A_P, as defined before where $t = \sum_{\{i_1, i_2, \dots i_n\} \subset I^n} \alpha_{i_1 i_2 \dots i_n}$. Each row has the form $(i_1\ i_2\ \cdots\ i_n)$ repeated for $\alpha_{i_1 i_2 \cdots i_n}$ times.
- we build the set $\Sigma(A_P)$ or $\Sigma'(A_P)$
- from each matrix $B \in \Sigma(A_P)$ or $B \in \Sigma'(A_P)$, using the inverse construction, we associate a new monomial $Q_B = \Pi_{\{j_1, j_2, \dots j_n\} \subset I^n} p_{j_1 j_2 \dots j_n}^{\alpha_{j_1 j_2 \dots j_n}}$

Definition 4.5.3. *Given a monomial $P = \Pi_{\{i_1, i_2, \dots i_n\} \subset I^n} p_{i_1 i_2 \dots i_n}^{\alpha_{i_1 i_2 \dots i_n}}$, we denote by $\overline{P = \Pi_{\{i_1, i_2, \dots i_n\} \subset I^n} p_{i_1 i_2 \dots i_n}^{\alpha_{i_1 i_2 \dots i_n}}}$ (respectively by $\underline{P = \Pi_{\{i_1, i_2, \dots i_n\} \subset I^n} p_{i_1 i_2 \dots i_n}^{\alpha_{i_1 i_2 \dots i_n}}}$) any element Q_B obtained with the previous procedure where $B \in \Sigma(A_P)$ (respectively $B \in \Sigma'(A_P)$).*

Example 4.5.4. Consider the first case of Example 4.5.1.

The matrices in $\Sigma'(A)$ are of the forms

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

The corresponding monomial are respectively

$$p_{123}^2 p_{331}$$
$$p_{123}p_{131}p_{323}$$
$$p_{123}p_{133}p_{321}$$
$$p_{121}p_{133}p_{323}$$

Then, the term $\underline{p_{123}p_{121}p_{333}}$ denotes any of the previous four monomials.

Let us notice that there are not elements $\overline{p_{123}p_{121}p_{333}}$ since $\Sigma(A) = \emptyset$. As a matter of fact no one of the previous matrices satisfies $\rho(\sigma(A)) = \rho(A)$.

To see an example with $\Sigma(A) \neq \emptyset$ consider the following matrix

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 1 \\ 3 & 3 & 3 \end{pmatrix}.$$

Then $\Sigma(A)$ consists of the following matrices

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

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

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

We conclude with two results for the case of three random variables.

Theorem 4.5.5 (C. Bocci, M. Benedettelli, F. Rapallo, [8])**.** *The following polynomials are invariants for* $\widetilde{M}_{1,3,1}$*:*

i) $P_{ijk}P_{i'j'k'} - \overline{P_{ijk}P_{i'j'k'}} = 0$ *with* $(i,j,k),(i',j',k') \in [n]^3 \setminus \Delta$*;*

ii) $P_{iii}P_{klm}P_{qrs} - \overline{P_{iii}P_{klm}P_{qrs}} = 0$ *with* $(k,l,m),(q,r,s) \in [n]^3 \setminus \Delta$*;*

iii) $P_{ijk}P_{i'j'k'}P_{i''j''k''} - \overline{P_{ijk}P_{i'j'k'}P_{i''j''k''}} = 0$ *with* $(i,j,k),(i',j',k'),(i'',j'',k'') \in [n]^3 \setminus \Delta$*;*

iv) $P_{iii}P_{klm}P_{qrs}P_{tnp} - \overline{P_{iii}P_{klm}P_{qrs}P_{tnp}} = 0$ *with* $(k,l,m),(q,r,s),(t,n,p) \in [n]^3 \setminus \Delta$*;*

v) $P_{iii}P_{klm}P_{qrs}P_{tnp}P_{zvw} - \overline{P_{iii}P_{klm}P_{qrs}P_{tnp}P_{zvw}} = 0$ *with* $(k,l,m),(q,r,s),(t,n,p),(z,v,w) \in [n]^3 \setminus \Delta$*;*

Theorem 4.5.6 (C. Bocci, M. Benedettelli, F. Rapallo, [8])**.** *The following polynomials are invariants for* $\widetilde{M}_{1,3,1}$*:*

i) $P_{ijk}P_{i'j'k'} - \underline{P_{ijk}P_{i'j'k'}} = 0$ *with* $(i,j,k),(i',j',k') \in [n]^3 \setminus \Delta$*;*

ii) $P_{iii}P_{klm} - P_{jjj}P_{klm} + \underline{P_{jjj}P_{klm}} - \underline{P_{iii}P_{klm}} = 0$ *with* $(k,l,m) \in [n]^3 \setminus \Delta$ *and at least two indices among* k,l *and* m *are different from* i *and* j*;*

iii) $P_{ijk}P_{lmn}P_{qrs} - \underline{P_{ijk}P_{lmn}P_{qrs}} = 0$ *with* $(i,j,k),(l,m,n),(q,r,s) \in [n]^3 \setminus \Delta$*;*

iv) $P_{iii}P_{klm}P_{k'l'm'} - P_{jjj}P_{klm}P_{k'l'm'} - \underline{P_{iii}P_{klm}P_{k'l'm'}} + \underline{P_{jjj}P_{klm}P_{k'l'm'}} = 0$ *e con* $(k,l,m),(k',l',m') \in [n]^3 \setminus \Delta$ *e with at least two indices among* k,l,m *and among* k',l',m' *different from* i *and* j*;*

(v) $P_{iii}P_{jjj}P_{qrs} - P_{iii}P_{kkk}P_{qrs} + P_{jjj}P_{kkk}P_{qrs} + P_{jjj}\underline{P_{jjj}P_{qrs}} - P_{iii}\underline{P_{jjj}P_{qrs}} + \underline{P_{iii}P_{kkk}P_{qrs}} - \underline{P_{jjj}P_{kkk}P_{qrs}} - P_{jjj}^2P_{qrs}$*;*

(vi) $P_{iii}P_{jjj}^2 - P_{iii}^2P_{jjj} + P_{iii}^2P_{kkk} - P_{iii}P_{kkk}^2 + P_{jjj}P_{kkk}^2 - P_{jjj}^2P_{kkk} + P_{iii}\underline{P_{iii}P_{jjj}} - P_{jjj}\underline{P_{iii}P_{jjj}} + P_{kkk}\underline{P_{iii}P_{kkk}} - P_{iii}\underline{P_{iii}P_{kkk}} + P_{jjj}\underline{P_{jjj}P_{kkk}} - P_{kkk}\underline{P_{jjj}P_{kkk}}$

(vii) $P_{iii}P_{klm}P_{k'l'm'}P_{k''l''m''} - P_{jjj}P_{klm}P_{k'l'm'}P_{k''l''m''} - \underline{P_{iii}P_{klm}P_{k'l'm'}P_{k''l''m''}} + \underline{P_{jjj}P_{klm}P_{k'l'm'}P_{k''l''m''}}$

viii) $P_{iii}P_{jjj}P_{klm} - P_{jjj}^2P_{klm} + P_{jjj}\underline{P_{jjj}P_{klm}} - P_{nnn}\underline{P_{iii}P_{klm}} - \underline{P_{iii}P_{jjj}P_{klm}} + \underline{P_{nnn}P_{iii}P_{klm}}$

Appendix A.

A.1 Topics on Commutative Algebra

A.1.1 Rings and ideals

A **ring** R is an abelian group $(R, +)$ with a multiplication operation $(a, b) \to ab$ and an identity element 1, satisfying, for all $a, b, c \in R$:

1) $a(bc) = (ab)c$ (associativity);

2) $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$ (distributivity);

3) $1a = a1 = a$ (identity).

A ring R is **commutative** if, moreover, $ab = ba$ for all $a, b \in R$. From now we will consider only commutative rings. An **invertible element** in a ring R is an element u such that there is an element $v \in R$ with $uv = 1$. It is a simple exercise to prove that such v is unique. We denote it by u^{-1} and it is called the **inverse** of u. A **field** is a ring in which every nonzero element is invertible. $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are fields.

Definition A.1.1. *An* **ideal** *in a commutative ring R is an additive subgroup I such that if $r \in R$ and $s \in I$, then $rs \in I$.*

An ideal I is said to be generated by a subset $S \subset R$ if every element $t \subset I$ can be written in the form

$$t = \sum_i r_i s_i \qquad ri \in R \text{ and } s_i \in S.$$

We shall write $\langle S \rangle$ for the ideal generated by a subset $S \subset R$; if S consists of finitely many elements $s_1, \ldots, s_n$ then we usually write $\langle s_1, \ldots, s_n \rangle$ in place of (S). By convention, the ideal generated by the empty set is (0). An ideal is principal if it can be generated by one element. An ideal I of a commutative ring R is **prime** if $I \neq R$ (we usually say that I is a proper ideal in this case) and if $f, g \in R$ and $fg \in I$ implies $f \in I$ or $g \in I$. The ring R is called a **domain** if (0) is prime. A maximal ideal of R is a proper ideal P not contained in any other proper ideal. If $P \subset R$ is a maximal ideal, then R/P is a field, so P is prime.

A ring homomorphism, or ring map, from a ring R to a ring S is a homomorphism of abelian groups that preserves multiplication and takes the identity element of R to the identity element of S.

Definition A.1.2. *A ring R is* **Noetherian** *if every ideal I of R is finitely generated, that is, there are element $f_1, \ldots, f_t \in R$ such that $I = \langle f_1, \ldots, f_t \rangle$.*

For example, any field is Noetherian (the only ideals are 0 and the whole field) and the ring $\mathbb{Z}$ of integers is Noetherian (each ideal is generated by a single integer, the greatest common divisor of the elements of the ideal).

A.1.2 Polynomial rings

If k is a commutative ring, then a **polynomial ring** over k in n variables $x_1, \ldots, x_n$ is denoted $k[x_1, \ldots, x_n]$. The elements of k are generally referred to as scalars. A **monomial** is a product of variables; its **degree** is the number of these factors (counting repeats) so that, for example, $x_1 x_3^2 x_4^2$ has degree 5. By convention the elements in k are seen as monomials of degree 0. A **term** is a scalar times a monomial. Every **polynomial** can be written uniquely as a finite sum of nonzero terms. If the monomials in the terms of a polynomial f all have the same degree (or if $f = 0$), then f is said to be **homogeneous**. Hilbert originally showed that a polynomial ring in n variables over a field or over the ring of integers is Noetherian.

Theorem A.1.3 (Hilbert Basis Theorem)**.** *If a ring R is Noetherian, then also the polynomial ring R[x] is Noetherian.*

A.1.3 Monomial orderings and Gröbner basis

The ring $k[x_1, \ldots, x_n]$ is an infinite-dimensional k-vector space, and it comes with a distinguished basis which is given by the set of monomials

$$x_1^{a_1} \cdot x^{a_2} \cdots x_n^{a_n}$$

where the a_i's run over $\mathbb{N}$. If $\mathbf{a} = (a_1, \ldots, a_n)$, we denote $x_1^{a_1} \cdot x^{a_2} \cdots x_n^{a_n}$ by $\mathbf{x^a}$. In order to write down a polynomial in R, it is convenient to fix a monomial order $<$. By this we mean a total order on the set of monomials which satisfies

$$1 < \mathbf{x^a}$$

and

$$\mathbf{x^a} < \mathbf{x^b} \Rightarrow \mathbf{x^{a+c}} < \mathbf{x^{b+c}} \text{ for all } \mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{N}^n.$$

We consider now a particular class of ideals, i.e the case of ideals generated by monomials. Such ideals are called monomial ideals. A monomial m lies in a given monomial ideal $M = (\mathbf{x^a}, \mathbf{x^b}, \mathbf{x^c}, \ldots)$ if and only if one of the generators of M divides m. By Dickson's Lemma we obtain the following

Lemma A.1.4. *Every monomial ideal M in R is finitely generated.*

Fix a monomial order $<$. Every polynomial $f \in R$ has a unique $(<)-$largest monomial $\mathbf{x}^{\mathbf{a}}$ which appears in f with non-zero coefficient. This monomial is called the **initial monomial** of f and it is denoted $\mathrm{In}_<(f)$. If I is any ideal in R then its **initial ideal** is the monomial ideal

$$\mathrm{In}_<(I) = \{\mathrm{in}_<(f) : f \in I\}.$$

By Lemma A.1.4, this ideal is finitely generated. Hence there exists a finite subset G of I such that

$$\mathrm{In}_<(I) = \{\mathrm{In}_<(g) : g \in G\}.$$

A subset G with this property is a **Gröbner basis** of I with respect to $<$. From Lemma A.1.4 one can derive the following result.

Lemma A.1.5. *Every Gröbner basis of I is a generating set of I.*

The Gröbner basis G is **reduced** if, for each $g \in G$, the initial monomial of g is a minimal generator of $\mathrm{In}_<(I)$ and none of the other monomials of g lies in $\mathrm{In}_<(I)$. The reduced Gröbner basis of I is unique when $<$ is fixed. Many computer algebra systems compute the reduced Gröbner basis G from any given generating set of the ideal. Once G is known, we can read off the monomial ideal $\mathrm{In}_<(I)$ and from this many invariants of I can be determined.

A.1.4 Elimination Theory

Elimination Theory is a systematic method do eliminate variables in a system of polynomial equations.

Definition A.1.6. *Given $I = \langle f_1, \ldots, f_t\rangle \subset k[x_1, \ldots, x_n]$, the $l-$th elimination ideal of I_l is the ideal in $k[x_{l+1}, \ldots, x_n]$ defined as*

$$I_l = I \cap k[x_{l+1}, \ldots, x_n].$$

It is easy to prove that I_l is an ideal in $k[x_{l+1}, \ldots, x_n]$. Obviously the ideal I_0 coincide with I. It is important to remark that different monomial ordering will produce different elimination ideals.

Thus, the elimination of $x_1, \ldots, x_l$ means to find the non-zero polynomials contained in the $l-$th elimination ideal. This can be done easily by Gröbner basis (once the monomial ordering is fixed).

Theorem A.1.7 (of elimination)**.** *Let $I \subset k[x_1, \ldots, x_n]$ be an ideal and G a Gröbner basis for I with respect to the lexicographical ordering with $x_1 > x_2 > \cdots > x_n$. Then, for any $0 \leq l \leq n$, the set*

$$G_l = G \cap k[x_{l+1}, \ldots, x_n]$$

is a Gröbner basis for the l'$l-$th elimination ideal I_l.

Example A.1.8. Consider an parametric algebraic statistical model

$$\begin{array}{rccc} f: & \Theta \subseteq \mathbb{R}^d & \to & \mathbb{R}^m \\ & \theta = (\theta_1, \ldots, \theta_d) & \mapsto & \Big(f_1(\theta), \ldots, f_m(\theta)\Big) \end{array}$$

If we want to find the invariant ideal of $\mathcal{M}_f$ we can use Elimination Theory in the following way: we define the ideal $I = \langle p_1 - f_1(\theta_1, \ldots, \theta_d), \ldots, p_m - f_m(\theta_1, \ldots, \theta_d)\rangle$ in $\mathbb{R}[\theta_1, \ldots, \theta_d, p_1, \ldots, p_m]$ and we eliminate, from I, the variables $\theta_1, \ldots, \theta_d$. The obtain elimination ideal is the invariant ideal of $\mathcal{M}_f$.

A.2 Topics on Algebraic Geometry

A.2.1 Affine geometry

Let k be a fixed algebraically closed field. We define **affine n-space** over k, denoted $\mathbb{A}^n_k$ or simply $\mathbb{A}^n_k$, to be the set of all $n-$tuples of elements of k. An element $P \in A^n$ will be called a point, and if $P = (a_1, \ldots, a_n)$ with $a_i \in k$, then the a_i will be called the coordinates of P.

Let $R = k[x_1, \ldots, x_n]$ be the polynomial ring in n variables over k. We will interpret the elements of A as functions from the affine n-space to k, by defining

$$f(P) = f(a_1, \ldots, a_n)$$

i.e, by making the substitution $a_i \to x_i$, where $f \in R$ and $P \in A^n$, $P = (a_1, \ldots, a_n)$. Thus if $f \in R$ is a polynomial, we can talk about the **zero set** of f, namely

$$Z(f) = \{P \in A^n : f(P) = 0\}.$$

More generally, if T is any subset of R, we define the zero set of T to be the common zeros of all the elements of T, namely

$$Z(T) = \{P \in A^n : f(P) = 0 \text{ for all } T\}.$$

Clearly if I is the ideal of R generated by T, then $Z(T) = Z(I)$. Furthermore, since R is a noetherian ring, by Theorem A.1.3, any ideal I has a finite set of generators $f_1, \ldots, f_r$. Thus $Z(T)$ can be expressed as the common zeros of the finite set of polynomials $f_1, \ldots, f_r$.

Definition A.2.1. *A subset Y of $\mathbb{A}^n$ is an algebraic set if there exists a subset $T \subset R$ such that $Y = Z(T)$.*

It is possible to define a topology on $\mathbb{A}^n$ by taking the open subsets to be the complements of the algebraic sets ([36] Proposition 1.1). This topology is called the **Zarisky topology**. We recall that a nonempty subset Y of a topological space X is **irreducible** if it cannot be expressed as the union $Y = Y_1 \cup Y_2$ of two proper subsets, each one of which is closed in Y

Definition A.2.2. *An* **affine algebraic variety** *(or simply* **affine variety***) is an irreducible closed subset of* $\mathbb{A}^n$ *(with the induced topology). An open subset of an affine variety is a* **quasi-affine variety**.

We pass now to explore the relationship between subsets of $\mathbb{A}^n$ and ideals in R more deeply. For any subset $Y \subset \mathbb{A}^n$, let us define the ideal of Y in R by

$$I(Y) = \{f \in R : f(P) = 0 \text{ for all } P \in Y\}.$$

Thus we have a way to obtain ideals of R starting from algebraic sets in $\mathbb{A}^n$, and, viceversa, algebraic sets in $\mathbb{A}^n$ starting from ideals of R. In particular one has

Proposition A.2.3.

(a) If $T_1 \subseteq T_2$ are subsets of R, then $Z(T_1) \supseteq Z(T_2)$.

(b) if $Y_1 \subseteq Y_2$ are subsets of $\mathbb{A}^n$, then $I(Y_1) \supseteq I(Y_2)$.

(c) For any two subsets Y_1, Y_2 of $\mathbb{A}^n$, we have $I(Y_1, \cup Y_2) = I(Y_1) \cap I(Y_2)$.

(d) For any ideal $a \subset R$, $I(Z(a)) = \sqrt{a}$, the radical of a (Hilbert's Nullstellensatz).

(e) For any subset $Y \subseteq A^n$, $Z(I(Y)) = \overline{Y}$, the closure of Y.

Proof. See [36], Proposition 1.2. □

Thus we finally state the following

Proposition A.2.4. *There is a one-to-one inclusion-reversing correspondence between algebraic sets in* $\mathbb{A}^n$ *and radical ideals (i.e., ideals which are equal to their own radical) in R, given by $Y \to I(Y)$ and $a \to Z(a)$. Furthermore, an algebraic set is irreducible if and only if its ideal is a prime ideal.*

Proof. [36] Corollary 1.4. □

Example A.2.5. Let f be an irreducible polynomial in $A = k[x, y]$. Then f generates a prime ideal in A, since A is a unique factorization domain ([26]) so the zero set $Y = Z(f)$ is irreducible. We call it the **affine curve** defined by the equation $f = 0$. More generally, if f is an irreducible polynomial in $A = k[x_1, \ldots, x_n]$, we obtain an affine variety $Y = Z(f)$, which is called a **surface** if $n = 3$, or a **hypersurface** if $n > 3$.

Remark A.2.6. Let T and I be respectively a set and an ideal of polynomials in R. T defines a variety V **set–theoretically** if V is the zero set of T, i.e. $V = Z(T)$. Instead I defines V **ideal-theoretically** if $I = I(V)$. In general one has $T \subset I(Z(T))$.

A.2.2 Projective geometry

To define projective varieties, we proceed in a manner analogous to the definition of affine varieties, except that we work in projective space. Let k be an algebraically closed field.The **projective $n-$space** over k, denoted $\mathbb{P}^n_k$, or simply $\mathbb{P}^n$, is defined as the set of equivalence classes of $(n+1)-$tuples $[x_0, \ldots, x_n]$ of elements of k, not all zero, under the equivalence relation given by $[x_0, \ldots, x_n] \sim [\lambda x_0, \ldots, \lambda x_n]$ for all $\lambda \in k \setminus \{0\}$. Equivalently we can say that $\mathbb{P}^n$, as a set, is the quotient of the set $\mathbb{A}^{n+1} \setminus \{0, \ldots, 0\}$ under the equivalence relation which identifies points lying on the same line through the origin. If $P = [x_0, \ldots, x_n]$ is a point in $\mathbb{P}^n$, then any $(n+1)-$tuple $[y_0, \ldots, y_n]$ in the equivalence class P is called a **set of homogeneous coordinates** for P. Let S be the polynomial ring $k[x_0, \ldots, x_n]$. If $f \in S$ is a polynomial, we cannot use it to define a function on $\mathbb{P}^n$ because of the nonuniqueness of the homogeneous coordinates. However, if f is a homogeneous polynomial of degree d, then

$$f(\lambda a_0, \ldots, \lambda a_n) = \lambda^d f(a_0, \ldots, a_n)$$

so that the property of f being zero or not depends only on the equivalence class of $[a_0, \ldots, a_n]$. Thus we can talk about the zeros of a homogeneous polynomial, namely

$$Z(f) = \{P \in \mathbb{P}^n : f(P) = 0\}.$$

Hence in the projective case we are interested only in polynomials f which are homogeneous. An ideal $I \subset S$ is homogeneous if and only if it can be generated by homogeneous elements. The sum, product, intersection, and radical of homogeneous ideals are homogeneous. As in the affine case, we can define $Z(T)$ and $Z(I)$ where T and I are respectively any set of homogeneous elements of S and a homogeneous ideal of S.

Definition A.2.7. *A subset Y of $\mathbb{P}^n$ is an algebraic set if there exists a set T of homogeneous elements of S such that $Y = Z(T)$.*

Again, we can define the Zarisky topology on $\mathbb{P}^n$ taking the algebraic sets as closed sets. Moreover, if $Y \subset \mathbb{P}^n$, then we can define

$$I(Y) = \{f \in S : f \text{ homogeneous and } f(p) = 0,\ \forall p \in Y\}$$

Definition A.2.8. *A* **projective variety** *is an irreducible algebraic set in $\mathbb{P}^n$.*

Remark A.2.9. We point out that the projective $n-$space has an open covering by affine $n-$spaces, and hence that every projective (respectively, quasi-projective) variety has an open covering by affine (respectively, quasi-affine) varieties.

Remark A.2.10. Also in the projective case, as in the affine one, we can speak of set-theoretically and ideal-theoretically description of a variety (Remark A.2.6).

A.2.3 Veronese embeddings

Fix non-negative integers n, d and $N = \binom{n+d}{d} - 1$. Let $\nu_{n,d} : \mathbb{P}^n \to \mathbb{P}^N$ be the map defined by sending $[x_0, \dots, x_n]$ to the set of monomials of degree d in $x_0, \dots, x_n$

$$[x_0^d, x_0^{d-1}x_1, x_0^{d-1}x_2, \dots, x_n^d]$$

ordered in lexicographic order. The map $\nu_{n,d}$ is well-defined and injective for all n and d and is called the $d-$**Veronese embedding** of $\mathbb{P}^n$. The image of $\nu_{n,d}$ is a subvariety of $\mathbb{P}^N$ called the d**-Veronese of** $\mathbb{P}^n$.

For example, the 3−Veronese embedding of $\mathbb{P}^1$ is the map

$$\begin{array}{ccc} \mathbb{P}^1 & \xrightarrow{\nu_{1,3}} & \mathbb{P}^3 \\ [x_0, x_1] & \mapsto & [z_0, z_1, z_2, z_3] \end{array}.$$

given by the parameterization

$$\begin{cases} z_0 = x_0^3 \\ z_1 = x_0^2 x_1 \\ z_2 = x_0 x_1^2 \\ z_3 = x_1^3 \end{cases}.$$

The 3-Veronese of $\mathbb{P}^1$, i.e. $\nu_{1,3}(\mathbb{P}^1)$, is a curve in $\mathbb{P}^3$ known as **twisted cubic curve**.

A.2.4 Segre embeddings

Let $\psi : \mathbb{P}^r \times \mathbb{P}^s \to \mathbb{P}^N$ be the map defined by sending the ordered pair $[a_0, \dots, a_r] \times [b_0, \dots, b_s]$ to $(\dots, a_i b_j, \dots)$ in lexicographic order, where $N = rs + r + s$. The map ψ is well-defined and injective and is called the **Segre embedding** of $\mathbb{P}^r \times \mathbb{P}^s$. The image of ψ is a subvariety of $\mathbb{P}^N$ called **Segre Variety**. The Segre embedding can be defined for an arbitrary number of factors

$$\psi : \mathbb{P}^{r_1} \times \cdots \times \mathbb{P}^{r_t} \to \mathbb{P}^N$$

where $N = \prod_{i=1}^{t}(r_i + 1) - 1$. For example, the Segre embedding of $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^2$ is the map

$$\begin{array}{ccccccc} \mathbb{P}^1 & \times & \mathbb{P}^1 & \times & \mathbb{P}^2 & \xrightarrow{\psi} & \mathbb{P}^{11} \\ [a_0, a_1] & & [b_0, b_1] & & [c_0, c_1, c_2] & \mapsto & [z_0, z_1, \dots, z_{11}] \end{array}.$$

given by the parameterization

$$\begin{cases} z_0 = a_0b_0c_0 \\ z_1 = a_0b_0c_1 \\ z_2 = a_0b_0c_2 \\ z_3 = a_0b_1c_0 \\ z_4 = a_0b_1c_1 \\ z_5 = a_0b_1c_2 \\ z_6 = a_1b_0c_0 \\ z_7 = a_1b_0c_1 \\ z_8 = a_1b_0c_2 \\ z_9 = a_1b_1c_0 \\ z_{10} = a_1b_1c_1 \\ z_{11} = a_1b_1c_2 \end{cases}.$$

A.2.5 Secant varieties

Due to their importance in our lecture, we collect here some known results on secant varieties.

We refer to [18], for details and proofs. We work over the complex field and we consider the projective space $\mathbb{P}^r = \mathbb{P}^r_{\mathbb{C}}$, equipped with the tautological line bundle $O_{\mathbb{P}^r}(1)$.

If Y is a subset of $\mathbb{P}^r$, we denote by $\langle Y \rangle$ the linear span of Y. We say that Y is **non–degenerate** if $\langle Y \rangle = \mathbb{P}^r$. A linear subspace of dimension n of $\mathbb{P}^r$ will be called a ***n*–subspace** of $\mathbb{P}^r$.

Let X be an irreducible projective variety X of dimension m, we denote by $S_k(X)$ the ***k*-th secant variety of** X, which is the Zariski closure of the set $\bigcup_{P_1,\dots,P_k \in X} \langle P_1, \dots, P_k \rangle$. In other words, $S_k(X)$ is the Zariski closure of the set of elements having X-rank equal to k.

$S_k(X)$ can be seen as the closure of the image, under the second projection, of the **abstract secant variety**, i.e. the incidence subvariety $AbS_k(X) \subset X^{(k)} \times \mathbb{P}^r$,

$$AbS_k(X) = \{((P_1, \dots, P_k), P) : P \in \langle P_1, \dots, P_k \rangle, \text{ independent } P_i\text{'s}\}.$$

Notice that $AbS_k(X)$ is *always* a variety of dimension $k(m+1)-1$. When $X \subset \mathbb{P}^r$ is reducible, the same definition of secant variety holds, except that we only consider linear spaces meeting every component of X. In particular, when X has k components, the secant variety coincides with the *join* of the components.

Definition A.2.11. *We say that X* **has k-th secant order** *μ if for a general point $P \in S_k(X)$, there are exactly μ unordered k-uples $P_1, \ldots, P_k$ of points of X such that $P \in \langle P_1, \ldots, P_k \rangle$.*

From the definition of secant varieties, it follows that:

$$S_{(k)}(X) := \dim(S_k(X)) \leq \min\{r, k(m+1) - 1\}. \tag{A.44}$$

The right hand side is called the **expected dimension** of $S_k(X)$.

Definition A.2.12. *We say that X is k***–defective** *when a strict inequality holds in (A.44).*

Let $X \subset \mathbb{P}^r$ be a variety. We denote by $\mathrm{Sing}(X)$ the Zariski-closed subset of singular points of X. Let $P \in X \setminus \mathrm{Sing}(X)$ be a smooth point. We denote by $T_{X,P}$ the embedded tangent space to X at P, which is a m–subspace of $\mathbb{P}^r$. More generally, if $P_1, \ldots, P_k$ are smooth points of X, we will set

$$T_{X,P_1,\ldots,P_k} = \langle \bigcup_{i=1}^{n} T_{X,P_i} \rangle.$$

The relations between secant varieties and tangent spaces to X are enlightened by the celebrated Terracini's Lemma:

Lemma A.2.13 ([53] or, for modern versions, [1], [23], [54])**.** *Given a general point $P \in S_k(X)$, lying in the subspace $\langle P_1, \ldots, P_k \rangle$ spanned by $k+1$ general points on X, then the tangent space $T_{S_k(X),P}$ to $S_k(X)$ at P is the span $T_{X,P_1,\ldots,P_k}$ of the tangent spaces $T_{X,P_1}, \ldots, T_{X,P_k}$.*

Using the correspondence between the abstract secant variety and $S_k(X)$, one obtains from Terracini's Lemma, a condition for the defectivity of X:

Theorem A.2.14 (L. Chiantini, C. Ciliberto [18], Theorem 2.5)**.** *Let $P_1, \ldots, P_k$ be general points of X. If H is a general hyperplane tangent to X at $P_1, \ldots, P_k$, we can consider the contact variety of H, i.e. the union Σ of the irreducible components of $\mathrm{Sing}(X \cap H)$. If X is k-defective, then Σ is positive dimensional.*

The previous Theorem suggests a refinement of the notion of defective variety.

Definition A.2.15. *An irreducible, non–degenerate variety $X \subset \mathbb{P}^r$ such that $S_{(k)}(X) < r$ is k–weakly defective if for $P_1, ..., P_k \in X$ general points, the general hyperplane H containing $T_{X,P_1,\ldots,P_k}$ is tangent to X along a variety $\Sigma(H)$ of positive dimension. $\Sigma(H)$ is called the k***–contact variety** *of H.*

It turns out that k-defective implies k–weakly defective, but the converse is false. We refer to [16] and [18] for a discussion on the subject.

The second cornerstone in our theory links k-defectivity and k-weakly defectivity with the existence of degenerate subvarieties, passing through k general points in X. Namely, if X is k-defective or k-weakly defective, then it turns out that the general contact variety is highly degenerate.

Theorem A.2.16 (L. Chiantini, C. Ciliberto, [18], Theorem 2.4 and Theorem 2.5)**.** *Assume* $k(m+1)-1 < r$*. If X is k-weakly defective, then a general contact variety* Σ *spans a linear space of dimension* $\leq k(n+1)-1$*, where* $n = \dim(\Sigma)$*. Moreover, X is k-defective if and only if* Σ *spans a space of dimension* $< k(n+1)-1$*.*

In conclusion, we obtain:

Corollary A.2.17. *Assume* $r > k(m+1)-1$*. Assume that for all* $n = 1, \ldots, m-1$*, there are no families of n-dimensional subvarieties of X, whose general element spans a linear space of dimension* $\leq k(n+1)-1$ *and passes through* $k+1$ *general points of X. Then X is not k-weakly defective..*

Bibliography

[1] B. Adlansdvik, *Joins and Higher secant varieties*, Math. Scand. **61** (1987), 213–222.

[2] A. Agresti, *Categorical Data Analysis*, Wiley, 2 ed., New York (2002).

[3] E.S. Allman, C. Matias, J.A. Rhodes, *Identifiability of parameters in latent structure models with many observed variables*. Ann. Statist. **37** (2009), 3099–3132.

[4] E. S. Allman, J. A. Rhodes, Phylogenetic invariants for the general Markov model of sequence mutation, Math. Biosci. **186** (2003), 113–144.

[5] E. S. Allman, J. A. Rhodes, *Phylogenetic invariants for stationary base composition*, J. Symbolic Comp. **41** no. 2 (2006), 138–150.

[6] E. S. Allman, J. A. Rhodes, *Phylogenetic ideals and varieties for the general Markov model*, Advances in Applied Mathematics **40** no. 2 (2008), 127–148.

[7] E. S. Allman, J. A. Rhodes, *Phylogenetics*, Lecture notes for AMS Short Course "Modeling and Simulation of Biological Networks", Chapter 2 in Proceedings of Symposia in Applied Mathematics, edited by R. Laubenbacher (2007).

[8] M. Benedettelli, C. Bocci, F. Rapallo, *Geometry of diagonal-effect models for probability tensors*, submitted.

[9] C. Bocci, *Topics on Phylogenetic Algebraic Geometry*, Expositiones Mathematicae, **25** no. 3 (2007), 235–259.

[10] C. Bocci, E. Carlini, F. Rapallo, *Geometry of diagonal-effect models for contingency tables*, In: Marlos A. G. Viana; Henry P. Wynn. Algebraic Methods in Statistics and Probability II. Vol. 516, AMS-CONM series - Urbana Volume, AMS (2010), 61–74.

[11] C. Bocci, L. Chiantini, *On the identifiability of binary Segre products*, J. Alg. Geom. **22** (2013), 1–11.

[12] C. Bocci, L. Chiantini, *Appendix to "On the identifiability of binary Segre products"*. Available at: `www.mat.unisi.it/personalpages/chiantini/p1app.pdf`.

[13] C. Bocci, L. Chiantini, G. Ottaviani, *Refined methods for the identifiability of tensors*, Ann. Mat. Pura Appl., to appear.

[14] M.V. Catalisano, A.V. Geramita, A. Gimigliano, *Secant varieties of* $\mathbb{P}^1 \times \ldots \times \mathbb{P}^1$ *(n-times) are not defective for* $n \geq 5$, J. Alg. Geom. **20** no. 2 (2011), 295–327.

[15] J. A. Cavender, J. Felsenstein, *Invariants of phylogenies in a simple case with discrete states*, J. of Class, **4** (1987), 57–71.

[16] L. Chiantini, C. Ciliberto, *Weakly defective varieties*, Trans. Amer. Math. Soc. **354** (2002), 151–178.

[17] L. Chiantini, C. Ciliberto, *On the classification of defective threefolds*, in C. Ciliberto, A. V. Geramita, B. Harbourne, R. M. Miro-Roig, and K. Ranestad (Ed.), Projective Varieties with Unexpected Properties: a volume in Memory of Giuseppe Veronese; proceedings of the international conference 'Varieties with Unexpected Properties', Siena, Italy, June 8-13, 2004, Walter de Gruyter (2005), 131–176.

[18] L. Chiantini, C. Ciliberto, *On the concept of k–secant order of a variety.* J. London Math. Soc. **73** (2006), 436–454.

[19] L. Chiantini, G. Ottaviani, *On generic identifiability of 3-tensors of small rank*, SIAM J. Matrix Anal. Appl., **33** no. 3, (2012), 1018–1037.

[20] CoCoATeam, *CoCoA: a system for doing Computations in Commutative Algebra*, Available at `http://cocoa.dima.unige.it` (2007).

[21] D/ Cox, J. Little, D. O'Shea, *Ideals, Varieties, and Algorithms*, Undergraduate Texts in Mathematics, Springer-Verlag, New York (1997).

[22] D. Cox, J. Little, D. O'Shea, *Using Algebraic Geometry* , Springer-Verlag, New York (1998).

[23] M. Dale, *Terracini's lemma and the secant variety of a curve*, Proc. London Math. Soc. **49** (1984), 329–339.

[24] L. De Lauthawer, *A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization*, SIAM J. Matrix Anal. Appl. **28** (2006), 642–666.

[25] P. Diaconis, B. Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Ann. Statist. **26** no. 1 (1998), 363–397.

[26] D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Graduate Texts in Mathematics, Springer-Verlag, New York (1994).

[27] N. Eriksson, *Toric ideal of homogeneous phylogenetic models*, ISSAC 2004, ACM, New York (2004), 149–154.

[28] N. Eriksson, K. Ranestad, B. Sturmfels, S. Sullivant, *Phylogenetic Algebraic Geometry*, in C. Ciliberto, A. V. Geramita, B. Harbourne, R. M. Miro-Roig, and K. Ranestad (Ed.), Projective Varieties with Unexpected Properties: a volume in Memory of Giuseppe Veronese; proceedings of the international conference 'Varieties with Unexpected Properties', Siena, Italy, June 8-13, 2004, Walter de Gruyter (2005), 177–197.

[29] S. N. Evans, T. P. Speed, *Invariants of some probability models used in phylogenetic inference*, Ann. Statitist. **21** no. 1 (1993), 355–377.

[30] J. Felsentein, *Inferring Phylogenies*, Sinauer Associates, Inc., Sunderland (2003).

[31] V. Ferretti, D. Sankoff, *The empirical discovery of phylogenetic invariants*, Adv. in Appl. Probab. **25** no. 2 (1993), 290–302.

[32] V. Ferretti, D. Sankoff, *Phylogenetic invariants for more general evolutionary models*, J. Theor. Biol. **173** (1995), 147–162.

[33] L. D. Garcia, M. Stillman, B. Sturmfels, *Algebraic Geometry of bayesian network*, J. Simbolic. Comp. **39** (2005), 331–355.

[34] D. Grayson, M. Stillman, *Macaulay 2, a software system for research in algebraic geometry*, Available at: `www.math.uiuc.edu/Macaulay2/`.

[35] J. Harris, *Algebraic Geometry: a first course.* Graduate Texts in Math., Springer-Verlag, New York (1992).

[36] R. Harthshorne, *Algebraic Geometry*, Graduate Texts in Math., Springer-Verlag, New York, (1977).

[37] R. Elmore, P. Hall, A. Neeman, *An application of classical invariant theory to identifiability in nonparametric mixtures.* Ann. Inst. Fourier **55** (2005), 1–28.

[38] M. D. Hendy, *The relationship between simple evolutionary tree models and observable sequence data*, Systematic Zoology **38** (1989), 310–321.

[39] M. D. Hendy, D. Penny, *A framework for the quantitative study of evolutionary trees*, Systematic Zoology, **38** (1989), 297–309.

[40] T. Kolda B. Bader, Tensor Decompositions and Applications, SIAM Review, **51** no. 3 (2009), 455–500.

[41] M. Kreuzer, L. Robbiano, *Computational Commutative Algebra 1*, Springer, Berlin (2000).

[42] J.B. Kruskal, *Three-way arrays: rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics*, Lin. Alg. Applic. **18** (1977), 95–138.

[43] J. A. Lake, *A rate independent technique for analysis of nucleic acid sequence: Evolutionary parsimony*, Mol. Bio. Evol., **4** (1987), 167–191.

[44] J. M. Landsberg, L. Manivel, *On the ideal of secant varieties of Segre varieties*, Found. Comput. Math., **4** no. 4 (2004), 397–422.

[45] L. Pachter, B. Sturmfels, *The Mathematics of phylogenomics*, SIAM Review **49** (2007), 3–31.

[46] L. Pachter, B. Sturmfels (Ed.), *Algebraic Statistics for Computational Biology*, Cambridge University Press (2005).

[47] G. Pistone, E. Riccomagno, H.P. Wynn, *Algebraic statistics: Computational commutative algebra in statistics*, Chapman&Hall/CRC, Boca Raton (2001).

[48] C. Semple, M. Steel, *Phylogenetics*, volume 24 of Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, Oxford (2003).

[49] R.D. Snee, *Graphical display of two-way contingency tables*, Amer. Statist. **38** (1974), 9–12.

[50] M. Steel, *Recovering a tree from the leaf colourations it generates under a Markov model*, Appl. Math. Letters **7** no. 2 (1994), 19–24.

[51] V. Strassen, *Rank and optimal computation of generic tensors*, Linear Algebra Appl. **52** (1983), 645–685.

[52] B. Sturmfels, S. Sullivant, *Toric ideals of phylogenetic invariants*, J. Comp. Biology, **12** no. 2 (2005), 204–228.

[53] A. Terracini, *Sulle V_k per cui la varietà degli S_h, $(h+1)$-seganti ha dimensione minore dell'ordinario*, Rend. Circ. Mat. Palermo **31** (1911), 392–396.

[54] F.L. Zak, *Tangents and secants of varieties.* AMS Bookstore pubblications, Transl. Math. Monog. **127** (1993).