Algebraic Statistics

Benjamin C. W. Brown
B.Brown@ed.ac.uk

31st October 2020

Contents I

planetmath.org

A *statistical model* is usually parameterised by a function, called a *parametrisation*

$$\Theta \to \mathcal{P}, \quad \text{given by} \quad \theta \mapsto P_\theta, \quad \text{so that} \quad \mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where Θ is the *parameter space.* Θ is usually a subset of $\mathbb{R}^n$.

McCullagh, 2002 [McC02]

This should be defined using category theory.

Note on References

General theory without particular references comes from either [PS05] or [DSS09].

"Does Watching Football on TV Cause Hair Loss?"

In a fictional study, 296 British subjects between the ages of 40 – 50 were asked about their hair loss and how much football they watch on television.

3×3 *Contingency Table*

We can represent the responses in a 3×3 *contingency table*:

	Hair Amount		
TV Time	lots	medium	balding
≥ 2h	51	45	33
$2-6$ h	28	30	29
≥ 6h	15	27	38

Based on the data, are these two random variables independent, or is there a correlation?

"Does Watching Football on TV Cause Hair Loss?"

$$M = \begin{bmatrix} 51 & 45 & 33 \\ 28 & 30 & 29 \\ 15 & 27 & 38 \end{bmatrix}$$

Null Hypothesis

H_0 : *Football on TV and Hair Loss are Independent.*

- Independence means that the (2×2)-minors of the data matrix M should vanish, but in fact they are all strictly quite positive, suggesting a positive correlation!
- For example, the top left (2×2)-minor equals: $51 \cdot 30 - 45 \cdot 28 = 1530 - 1260 = 270 \neq 0$.

- A better explanation for our data is obtained by identifying a certain *hidden variable*, which is the *gender identification* of the respondents:

$$M = M_m + M_f = \underbrace{\begin{bmatrix} 3 & 9 & 15 \\ 4 & 12 & 20 \\ 7 & 21 & 35 \end{bmatrix}}_{\text{126 male}} + \underbrace{\begin{bmatrix} 48 & 36 & 18 \\ 24 & 18 & 9 \\ 8 & 6 & 3 \end{bmatrix}}_{\text{170 female}}.$$

Alternative Hypothesis

Instead, we have *conditional independence*:

H_0 : *Football on TV & Hair Loss are Independent* given *Gender.*

- Classify using two criteria with r and c levels, yielding two random variables X and Y.
- Note outcomes as $[r] := \{1, \ldots, r\}$, and $[c] := \{1, \ldots, c\}$.

All information about X and Y is contained in the *joint probabilities*:

$$p_{ij} = P(X = i; Y = j), \quad i \in [r], \ j \in [c].$$

- These in turn determine the *marginal probabilities*:

$$p_{i+} := \sum_{j=1}^{c} p_{ij} = P(X = i), \quad i \in [r],$$

$$p_{+j} := \sum_{i=1}^{r} p_{ij} = P(Y = j), \quad j \in [c].$$

Definition

Two random variables X and Y are *independent* if the joint probabilities factor as $p_{ij} = p_{i+} \cdot p_{+j}$, for all $i \in [r]$ and $j \in [c]$. Denote independence of X and Y by $X \perp\!\!\!\perp Y$.

Proposition

Two random variables X and Y are independent if and only if the $(r \times c)$-matrix, $p = (p_{ij})$, has rank one.

For a (2×2)-table, we thus have:

$$\begin{array}{c|cc} & P(Y=1) & P(Y=2) \\ \hline P(X=1) & p_{11} & p_{12} \\ P(X=2) & p_{21} & p_{22} \end{array} \quad \overset{X \perp\!\!\!\perp Y}{\rightsquigarrow} \quad p_{11}p_{22} = p_{12}p_{21}.$$

Suppose now we select n cases, giving rise to n independent pairs of discrete random variables:

$$\begin{pmatrix} X^{(1)} \\ Y^{(1)} \end{pmatrix}, \begin{pmatrix} X^{(2)} \\ Y^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} X^{(n)} \\ Y^{(n)} \end{pmatrix},$$

all drawn from the same distribution, i.e.:

$$P(X^{(k)} = i; Y^{(k)} = j) = p_{ij}, \quad \text{for all } i \in [r],\ j \in [c],\ k \in [n].$$

Probability Simplices

Joint probability matrix $p = (p_{ij})$ is an *unknown* element of the $(rc-1)$-dimensional *probability simplex*:

$$\Delta_{rc-1} = \left\{ q \in \mathbb{R}^{r\times c} \,\middle|\, q_{ij} \geq 0,\ \text{for all } i, j,\ \text{and } \sum_{i=1}^{r}\sum_{j=1}^{c} q_{ij} = 1 \right\}.$$

Definitions

A *statistical model* $\mathcal{M}$ is a subset of Δ_{rc-1}. It represents the set of all candidates for the unknown distribution p.
The *independence model* for X and Y is the set

$$\mathcal{M}_{X \perp\!\!\!\perp Y} := \{\, p \in \Delta_{rc-1} \mid \operatorname{rank}(p) = 1 \,\}\,.$$

$\mathcal{M}_{X \perp\!\!\!\perp Y}$ is the intersection of Δ_{rc-1} and the set of all matrices $p = (p_{ij})$ such that

$$p_{ij}p_{kl} - p_{il}p_{jk} = 0, \quad (1 \leq i < k \leq r, \text{ and } 1 \leq j < l \leq c).$$

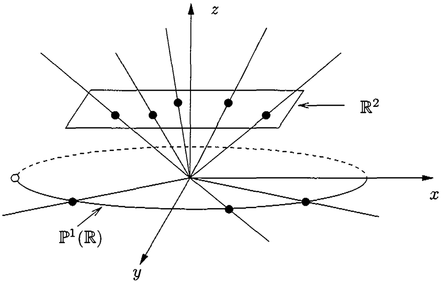
These are called *Segre varieties* in algebraic geometry.

Projective Space

Playing field is *n-dimensional projective space*, $\mathbb{P}^n$:

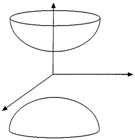
$$\mathbb{P}^n := \{(z_0, \ldots, z_n) \in \mathbb{C}^n\}/(\mathbf{x} \sim \lambda \cdot \mathbf{y}), \quad \lambda \neq 0,$$

that is, its elements consists of *lines through the origin* in $\mathbb{C}^n$.

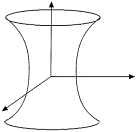


Varieties

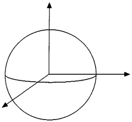
Varieties are the objects studied in algebraic geometry, determined by the *vanishing set*[1] $V(-)$, for a system of polynomials.



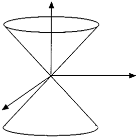
(a) $V(x^2 + y^2 - z^2 + 1)$



(b) $V(x^2 + y^2 - z^2 - 1)$



(c) $V(x^2 + y^2 + z^2 - 1)$



(d) $V(x^2 + y^2 - z^2)$

[1]from ‘*Verschwindungsmenge*’

Segre varieties come from $\sigma : \mathbb{P}^n \times \mathbb{P}^m \to \mathbb{P}^{(n+1)(m+1)-1}$, that sends $([X], [Y])$ to the pairwise products of their components:

$$\sigma : ([X_1, \dots, X_{n+1}], [Y_1, \dots, Y_{m+1}]) \mapsto [\dots, X_i Y_j, \dots].$$

Example (Segre quadric surface)

$\sigma : \mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^3$, $([X_1, X_2], [Y_1, Y_2]) \mapsto [X_1 Y_1, X_1 Y_2, X_2 Y_1, X_2 Y_2]$.

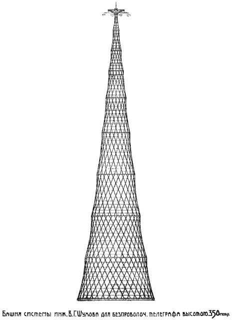
Set $[X_1 Y_1, X_1 Y_2, X_2 Y_1, X_2 Y_2] =: [p_{11}, p_{12}, p_{21}, p_{22}]$, then:

$$\rightsquigarrow \det \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = 0 \iff \operatorname{rank} \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \leq 1.$$

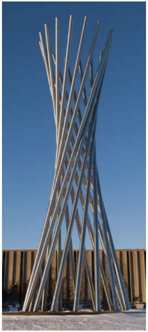
Rulings

The Segre quadric surface has two families of lines in it, called *rulings*. These are the images of $\sigma(\mathbb{P}^1 \times \{\mathrm{pt}\})$ and $\sigma(\{\mathrm{pt}\} \times \mathbb{P}^1)$.

Shukhov Tower, Nizhny Novgorod



Shukhov Tower, Moscow



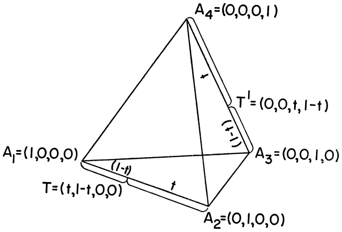
Tractricious, Fermilab

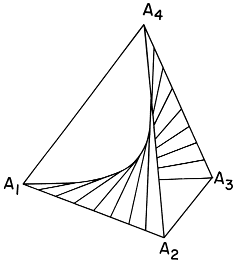
- Let $\Delta_3 \subset \mathbb{R}^4$, with vertices $A_i = e_i$, and let $p = (p_{ij}) \in \Delta_3$ be

$$p_{ij} = (p_{11}, p_{12}, p_{21}, p_{22}) = \begin{array}{|c|c|} \hline p_{11} & p_{12} \\ \hline p_{21} & p_{22} \\ \hline \end{array}$$

- Has been shown that the two rulings are given by [FG70]:

$$\begin{array}{|c|c|} \hline st & s(1-t) \\ \hline t(1-s) & (1-s)(1-t) \\ \hline \end{array} \qquad (0 \le s, t \le 1).$$





Hidden Variables Omake r: Known Unknowns

- Suppose $\mathcal{P} \subset \Delta_{r-1}$ is a model for a random variable X with state space $[r]$.
- Assume Y is a *hidden* or *latent* random variable, with state space $[s]$; for each $j \in [s]$, the conditional distribution of X given $Y = j$ is $p^{(j)} \in \mathcal{P}$.
- Y also has some probability distribution $\pi \in \Delta_{s-1}$.

So the joint distribution of Y and X is given by the formula

$$P(Y = j; X = i) = \pi_j \cdot p_i^{(j)}.$$

Donald Rumsfeld, 2002

"[T]here are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know."

- But as Y is hidden, we can only observe the marginal distribution of X, that is

$$P(X = i) = \sum_{j=1}^{s} \pi_j \cdot p_i^{(j)}.$$

- In other words, the marginal distribution of X is the convex combination of the s distributions $p^{(1)}, \dots, p^{(s)}$, with weights given by π.

Definition [DSS09]

Let $\mathcal{P} \subset \Delta_{r-1}$ be a statistical model. The *s-th mixture model* is

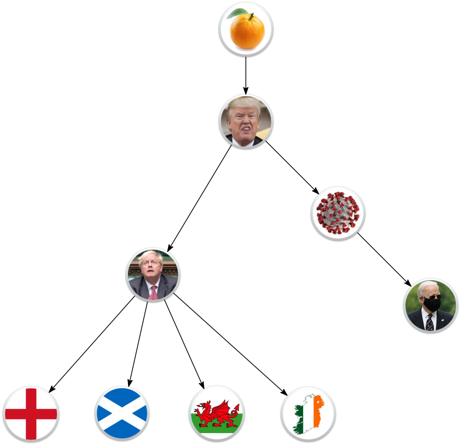
$$\mathrm{Mixt}^s(\mathcal{P}) := \left\{ \sum_{j=1}^{s} \pi_j \cdot p^{(j)} \;\middle|\; \pi \in \Delta_{s-1},\ p^{(j)} \in \mathcal{P},\ \text{for all } j \right\}.$$

Mixture Models

- Mixture models provide ways to build complex models out of simpler ones.
- Basic assumption is that the underlying population to be modelled can be split into s disjoint sub-populations.
- Restricted to each sub-population, the observable X follows a probability distribution from the simple model $\mathcal{P}$.
- After marginalisation though, the structure becomes significantly more complex as it is now a convex combination of these simple distributions.

Phylogenetic Trees

- Introduce *phylogenetic trees*; describe the descent of species from a common ancestor:



- Sequence of DNA molecules in a genome is represented as a sequence of letters from the four letter alphabet $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.
- *Fix for now* an ancestral nucleotide $\mathsf{Y} \in \Sigma$; we assume that the following evolution events occur independently [All07]:

$$\mathsf{Y} \xmapsto{\pi_{\mathsf{Y}} \cdot p_{\mathsf{A}}^{(\mathsf{Y})}} \mathsf{A}, \quad \mathsf{Y} \xmapsto{\pi_{\mathsf{Y}} \cdot p_{\mathsf{C}}^{(\mathsf{Y})}} \mathsf{C}, \quad \mathsf{Y} \xmapsto{\pi_{\mathsf{Y}} \cdot p_{\mathsf{G}}^{(\mathsf{Y})}} \mathsf{G}, \quad \mathsf{Y} \xmapsto{\pi_{\mathsf{Y}} \cdot p_{\mathsf{T}}^{(\mathsf{Y})}} \mathsf{T},$$

- So *given* Y, we have a joint distribution:

$$\pi_{\mathsf{Y}} \cdot [p_{\mathsf{A}}^{(\mathsf{Y})}, p_{\mathsf{C}}^{(\mathsf{Y})}, p_{\mathsf{G}}^{(\mathsf{Y})}, p_{\mathsf{T}}^{(\mathsf{Y})}] \in \Delta_3 = \Delta_{4-1}.$$

Example

- Y is a hidden variable though; could have been anything from $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.
- For *exactly one given choice* of Y, we had the distribution Δ_3; need to consider *all choices* of ancestral nucleotide Y.
- Hence we get the mixture model [All07]:

$$\mathrm{Mixt}^4(\Delta_3) = \left\{ \sum_{\mathsf{Y}\in\Sigma} \pi_{\mathsf{Y}} \cdot p^{(\mathsf{Y})} \;\middle|\; \pi \in \Delta_3,\ p^{(\mathsf{Y})} \in \mathcal{P} \subseteq \Delta_3,\ \text{for each } \mathsf{Y} \right\}.$$

Question?

What is the analogue for mixture models in algebraic statistics?

Answer!

Secant[2] varieties [DSS09]!

Definitions

- Consider two varieties $V, W \subseteq \mathbb{R}^k$. The *join* of V and W is the variety

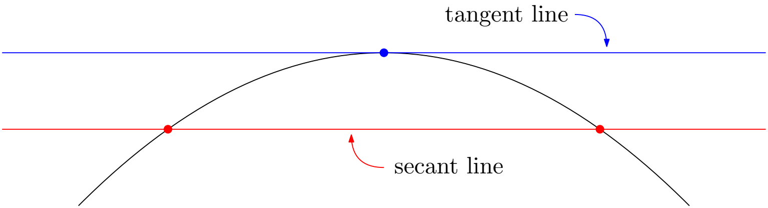
$$\mathcal{J}(V, W) := \{\lambda v + (1 - \lambda) w : v \in v, w \in W, \lambda \in [0, 1]\}.$$

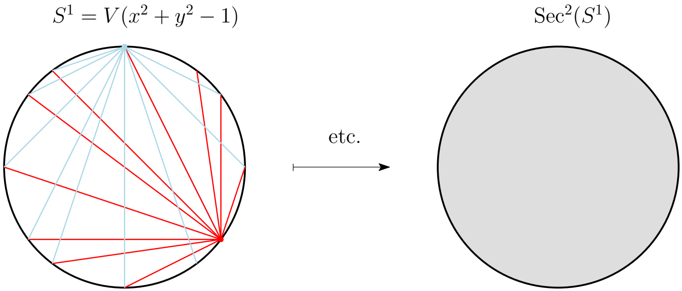
- If $V = W$, then this is the *secant variety* of V, denoted $\mathrm{Sec}^2(V) = \mathcal{J}(V, V)$. The *s-th higher secant variety* is:

$$\mathrm{Sec}^1(V) := V, \qquad \mathrm{Sec}^s(V) := \mathcal{J}(\mathrm{Sec}^{s-1}(V), V).$$

[2] from *secare*, "to cut" in Latin; *c.f. tangō*, "to touch".

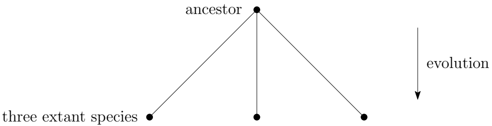
Secant Varieties





More Complicated Phylogenetic Trees

- Last example only had one extant species; what about if we had three extant species, all coming from the same ancestor?



- Now we have to consider: $\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$; or equivalently $\mathrm{Mixt}^4(\Delta_3 \times \Delta_3 \times \Delta_3)$ [All07].
- Finding the minimal set of polynomials defining $\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$ once gave rise to a very important application of algebraic statistics...

The *Salmon Problem*

Statement

Determine the ideal[3] *defining* $\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$, [All07].

Prize

- At an IMA workshop in 2007, Elizabeth Allman stated that she would personally catch and smoke copper river salmon from Alaska for whomever solved this problem.
- Solved in 2010 by Shmuel Friedland & Elizabeth Gross [FG12] (see [BO11] too for an in-depth discussion).

Solving this would then provide all polynomial invariants of the statistical model for any binary evolutionary tree, with any number of states [AR08]; [BO11].

[3] read this as “set of defining polynomials”.

Why $\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$ again?

- Three independent variables (nucleotides in extant species) $\rightsquigarrow$ three factors in product;
- Each independently assumes one value from $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ $\rightsquigarrow$ distribution is a point in $\mathbb{P}^3 = \mathbb{P}^{4-1}$;
- The ancestral nucleotide is unknown, but could assume any of the four values in Σ $\rightsquigarrow$ mix four such independence models;
- The model for the three observed nucleotides is therefore

$$\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3), \quad \textit{c.f.}, \quad \mathrm{Mixt}^4(\Delta_3 \times \Delta_3 \times \Delta_3).$$

A Dank Meme

Henri Poincaré

"[L]a mathématique est l'art de donner le même nom à des choses différentes," [Poi96].

H. Poincaré, 1887.

H. Poincaré, colourised.

The solution to the salmon conjecture is equivalent to [Stu09]:

- the mixture of four models for three independent variables;
- the fourth secant variety of the Segre variety $\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3$;
- the set of $(4 \times 4 \times 4)$-tables of tensor rank ≤ 4;
- the naive Bayes model with four classes and three features;
- the conditional independence model $[X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3 | Y]$;
- the general Markov model for the phylogenetic tree, $K_{1,3}$;
- the superposition of four pure states in a quantum system, [BH01]; [Hey06].

A 'Statistics to Algebraic Geometry' Lexicon, [PS05]

Statistics		Algebra/Geometry
independence	$\sim$	Segre variety
exponential family (log-linear models)	$\sim$	toric variety
curved exponential family	$\sim$	manifold
mixture model	$\sim$	secant variety
inference	$\sim$	tropicalisation
	$\vdots$	

Applications

We finish by mentioning that algebraic statistics thus has at least a few important & interesting applications:

- It can win you salmon;
- It can win you 100 Swiss francs[4] (CHF $100 \sim$ £85);
- One gets to learn lots of big words;
- It can provide one with a topic for an (excellent) colloquium talk;
- Algebraists & statisticians may try to talk to one other (this may be a con rather than a pro).

---

[4] The 100 *Swiss Francs Conjecture* was on maximising the likelihood function over the space of 4×4 stochastic matrices, of rank ≤ 2 [Stu09], and was solved in [ZJG11].

Questions?

References I

- Elizabeth Allman. *OPEN PROBLEM: Determine the ideal defining* $\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$. 2007. URL: `https://eallman.github.io/papers/salmonPrize.pdf`.
- Elizabeth S. Allman and John A. Rhodes. 'Phylogenetic ideals and varieties for the general Markov model'. In: *Adv. in Appl. Math.* 40.2 (2008).
- Dorje C. Brody and Lane P. Hughston. 'Geometric quantum mechanics'. In: *J. Geom. Phys.* 38.1 (2001).
- Daniel J. Bates and Luke Oeding. 'Toward a salmon conjecture'. In: *Exp. Math.* 20.3 (2011).

References II

- Mathias Drton, Bernd Sturmfels and Seth Sullivant. *Lectures on algebraic statistics.* Vol. 39. Oberwolfach Seminars. Birkhäuser Verlag, Basel, 2009.
- Shmuel Friedland and Elizabeth Gross. 'A proof of the set-theoretic version of the salmon conjecture'. In: *J. Algebra* 356 (2012), pp. 374–379.
- Stephen E. Fienberg and John P. Gilbert. 'The Geometry of a Two by Two Contingency Table'. In: *Journal of the American Statistical Association* 65.330 (1970).
- Hoshang Heydari. 'General pure multipartite entangled states and the Segre variety'. In: *J. Phys. A* 39.31 (2006).

References III

- Peter McCullagh. 'What is a statistical model?' In: *Ann. Statist.* 30.5 (2002).
- Henri Poincaré. *Science and method.* Key Texts: Classic Studies in the History of Ideas. Thoemmes Press, Bristol, 1996.
- Lior Pachter and Bernd Sturmfels, eds. *Algebraic statistics for computational biology*. Cambridge University Press, New York, 2005.
- Bernd Sturmfels. 'Open problems in algebraic statistics'. In: *Emerging applications of algebraic geometry*. Vol. 149. IMA Vol. Math. Appl. Springer, New York, 2009.

References IV

- Mingfu Zhu, Guangran Jiang and Shuhong Gao. 'Solving the 100 Swiss francs problem'. In: *Math. Comput. Sci.* 5.2 (2011).