

Summer School Problems

Seth Sullivant

Department of Mathematics and Society of Fellows, Harvard University

(Note that a (*) indicates a problem that I don't know the answer to and is a good problem for performing experiments.)

1 Lecture 1: Statistical Models are Algebraic Varieties

Problem 1.1. Prove that the joint distribution matrix

$$\frac{1}{8} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

is not in the mixture model $\text{Mixt}^3(\mathcal{M}_{X_1 \perp\!\!\!\perp X_2})$.

Problem 1.2. Prove that $\text{Sec}^2(\mathcal{M}_{X_1 \perp\!\!\!\perp X_2}) = \text{Mixt}^2(\mathcal{M}_{X_1 \perp\!\!\!\perp X_2})$ for all m_1 and m_2 .

Problem 1.3. Consider a 3-dimensional discrete random vector $X = (X_1, X_2, X_3)$.

1. What does $X_1 \perp\!\!\!\perp X_2, X_3$ mean in terms of the joint distribution (p_{ijk}) ? (Note that this reads “ X_1 is independent of X_2 and X_3 ”.) Give an algebraic characterization of all such joint distributions and determine the ideal $I(\mathcal{M}_{X_1 \perp\!\!\!\perp X_2, X_3})$.
2. What does $X_1 \perp\!\!\!\perp X_2 | X_3$ mean in terms of the joint distribution (p_{ijk}) ? (Note that this reads “ X_1 is independent of X_2 conditional on X_3 ”.) Give an algebraic characterization of all such joint distributions and determine the ideal $I(\mathcal{M}_{X_1 \perp\!\!\!\perp X_2 | X_3})$. What about for $X_1 \perp\!\!\!\perp X_3 | X_2$?
3. Do the two independence conditions $X_1 \perp\!\!\!\perp X_2 | X_3$ and $X_1 \perp\!\!\!\perp X_3 | X_2$ imply that $X_1 \perp\!\!\!\perp X_2, X_3$? How can you check this algebraically?

2 Lecture 2: Log-linear Models, Toric Varieties, and Their Markov Bases

Problem 2.1. Write down the matrix $A_{\Gamma, \mathbf{m}}$, and the monomial parametrization $\phi_{A_{\Gamma, \mathbf{m}}}$ in the case when:

1. $\Gamma = [12][13][24][34]$, and
2. $\Gamma = [12][13][234]$,

with $\mathbf{m} = (2, 2, 2, 2)$.

Problem 2.2. 1. Show that if $\Gamma^1 \subseteq \Gamma^2$ (that is, for each $S \in \Gamma^1$ there is a $T \in \Gamma^2$ such that $S \subseteq T$) then $I_{\Gamma^2, \mathbf{m}} \subseteq I_{\Gamma^1, \mathbf{m}}$.

2. Suppose a Markov basis for $A_{\Gamma^1, \mathbf{m}}$ is known and a Markov basis for $A_{\Gamma^2, \mathbf{m}}$ is too hard to compute (where $\Gamma^1 \subseteq \Gamma^2$). How might the Markov basis for $A_{\Gamma^1, \mathbf{m}}$ be used to perform conditional inference for the model $\mathcal{M}_{\Gamma^2, \mathbf{m}}$?

Problem 2.3. Give a combinatorial description of the quadratic binomials in $I_{A_{\Gamma, \mathbf{m}}}$ for any Γ . Under what circumstances does $I_{A_{\Gamma, \mathbf{m}}}$ not contain any quadrics?

Problem 2.4. A Markov basis \mathcal{F} for A is called a *minimal Markov basis* for A if there is no proper subset $\mathcal{F}' \subset \mathcal{F}$ that is also a Markov basis for A . Is there always a unique minimal Markov basis for A ? Give a proof or counterexample.

Problem 2.5. Suppose that $\mathbf{f} \in \ker_{\mathbb{Z}}(A)$ is in every Markov basis of A . Describe the fiber $A^{-1}[A\mathbf{f}^+]$ of such a move.

The *support* of a vector $\mathbf{u} \in \mathbb{N}^m$ is the set $\text{supp}(\mathbf{u}) = \{i \in [m] \mid u_i \neq 0\}$. Given A and \mathbf{b} define the simplicial complex $\Delta_{A, \mathbf{b}}$ on $[n]$ is generated by the supports of the \mathbf{u} in the fiber $A^{-1}[\mathbf{b}]$; that is,

$$\Delta_{A, \mathbf{b}} = \{S \subset \text{supp}(\mathbf{u}) \mid \mathbf{u} \in A^{-1}[\mathbf{b}]\}.$$

Problem 2.6. Prove that there exists $\mathbf{u}, \mathbf{v} \in A^{-1}[\mathbf{b}]$ such that $\mathbf{u} - \mathbf{v}$ belongs to some minimal Markov basis for A if and only if $\Delta_{A, \mathbf{b}}$ is disconnected.

3 Lecture 3: Algebraic Tools in Statistical Disclosure Limitation

Problem 3.1. Determine formulas for the sharp upper and lower bounds on a cell entry of a 2-way table with fixed row and column sums.

Problem 3.2. Generalizing the previous problem, determine formulas for the sharp upper and lower bounds on a cell entry of an n -way table, given all its 1-way margins. The associated simplicial complex is $\Gamma = [1][2] \cdots [n]$

Problem 3.3. Suppose that \mathbf{f} belongs to every Markov basis of A and that $f_1 > 0$. Show that $\text{Gap}(A, \mathbf{e}_1) \geq f_1 - 1$ where \mathbf{e}_1 is the first standard unit vector. (Hint: See Problem 2.5.)

Problem 3.4. Show that a lexicographic Gröbner basis for I_A can be used to solve integer programs with $\mathbf{c} = \mathbf{e}_1$. What integer programs does a reverse lexicographic Gröbner basis solve?

The shuttle algorithm is an easy-to-implement, iterative algorithm for approximating the integer upper and lower bounds on all entries of a vector \mathbf{u} given the margins $A\mathbf{u}$.

Algorithm 3.5. (Shuttle Algorithm)

- Input: A and \mathbf{b} .
- Output: Vectors L and U of approximate lower and upper bounds on the cell entries $A^{-1}[\mathbf{b}]$
- Set: $L = (0, 0, \dots, 0)$ and $U = (+\infty, +\infty, \dots, +\infty)$.
- Until convergence Do
 1. For $j = 1$ to m , Set

$$U_j = \min_{i: A_{ij} \neq 0} \left\lfloor \frac{b_i - \sum_{k \neq j} A_{ik} L_k}{A_{ij}} \right\rfloor$$

2. For $j = 1$ to m Set:

$$L_j = \max_{i: A_{ij} \neq 0} \left\lceil \frac{b_i - \sum_{k \neq j} A_{ik} U_k}{A_{ij}} \right\rceil.$$

Problem 3.6. Prove that the shuttle algorithm converges in finitely many steps to bounds on the entries of $A^{-1}[\mathbf{b}]$.

Problem 3.7. Suppose that A is a 0/1 matrix. Show that the bounds produced by the shuttle algorithm cannot be better than the bounds produced by linear programming.

Problem 3.8. (*) Suppose that A is a 0/1 matrix. What conditions on A (and, in particular, $\text{cone}(A)$) guarantee that the shuttle algorithm converges to the sharp integer bounds for all \mathbf{b} ?

4 Lecture 4: Maximum Likelihood Estimation

Problem 4.1. Let $A \in \mathbb{N}^{d \times m}$ and $\mathbf{h} \in \mathbb{R}_{>0}^m$ and let \mathbf{u} be a integer vector of data. Show that the maximum likelihood estimate $\hat{\mathbf{p}} \in \mathcal{M}_{A,\mathbf{h}}$ (if it exists) is the unique nonnegative root of the ideal

$$I_{A,\mathbf{h}} + \left\langle A\mathbf{p} - A \frac{\mathbf{u}}{\|\mathbf{u}\|_1} \right\rangle,$$

where the second ideal denotes a set of linear constraints on \mathbf{p} .

Problem 4.2. (*) Let A be a $2 \times m$ matrix with first row equal to $(1, 1, 1, \dots, 1)$ and $\mathbf{h} \in \mathbb{R}_{\geq 0}^m$. Compute some examples of the maximum likelihood degree of $\mathcal{M}_{A\mathbf{h}}$, with varying A and \mathbf{h} . What patterns do you notice? For fixed A how does the maximum likelihood degree depend on \mathbf{h} ?

Problem 4.3. (*) Let $\phi_1, \phi_2, \dots, \phi_m$ be polynomials each with fixed monomial support, but whose coefficients are considered as indeterminate. Any fixed value of the coefficients determines a rational map ϕ . What do you think the *maximum likelihood discriminant* of these polynomials means? Can you compute the ml-discriminant in any small cases? How does this relate to the preceding problem?

5 Lecture 5: Phylogenetic Algebraic Geometry

Problem 5.1. Consider the general Markov model on the three leaf tree $K_{1,3}$. Identify (the Zariski closure of) this phylogenetic model as a certain classical algebraic variety. What other possible statistical interpretations are there for this model, in terms of constructions we have already seen?

Problem 5.2. Given an unrooted tree T , a subforest F of T is a subgraph of T such that every leaf of every connection component of F is a leaf of T as well. If T is a trivalent tree with n leaves, how many distinct subforests of T are there? (Note: this is the number of distinct Fourier coordinates of the Jukes-Cantor DNA model on T .)

Problem 5.3. The Jukes-Cantor DNA model on an unrooted four leaf tree T has the following parametric description in its Fourier coordinates:

$$\begin{aligned} q_{00000} &= a_0 b_0 c_0 d_0 e_0, & q_{00011} &= a_0 b_0 c_0 d_1 e_1, & q_{11000} &= a_1 b_1 c_0 d_0 e_0, \\ q_{11011} &= a_1 b_1 c_0 d_1 e_1, & q_{10110} &= a_1 b_0 c_1 d_1 e_0, & q_{10101} &= a_1 b_0 c_1 d_0 e_1, \\ q_{00110} &= a_0 b_1 c_1 d_1 e_0, & q_{01101} &= a_0 b_1 c_1 d_0 e_1, & q_{11101} &= a_1 b_1 c_1 d_0 e_1, \\ q_{01111} &= a_0 b_1 c_1 d_1 e_1, & q_{10111} &= a_1 b_0 c_1 d_1 e_1, & q_{11110} &= a_1 b_1 c_1 d_1 e_0, \end{aligned}$$

$$q_{111111} = a_1 b_1 c_1 d_1 e_1.$$

The parameters $a_\bullet, b_\bullet, d_\bullet, e_\bullet$ correspond to the leaves of the tree and the parameters c_\bullet correspond to the internal edge.

1. Show that these 13 Fourier coordinates correspond to the 13 subforests of T .
2. What does it mean (in terms of the subforests) for a binomial $q^{\mathbf{u}} - q^{\mathbf{v}}$ to belong to the phylogenetic ideal I_T for this example?
3. Compute the generators of the phylogenetic ideal I_T .
4. (*) Compute generators of the ideal $I(\text{Mixt}^2(V_T))$ defining the mixture model for this phylogenetic model. What is the biological meaning (in terms of mutating DNA sequences) of such a mixture model?

Problem 5.4. Compute the ml-degree of some small phylogenetic models.

6 Lecture 6: Combinatorial Secant Varieties

Problem 6.1. Compute the secant ideal of the graph ideals $I(G)$ for the following graphs:

1. G such that $V(G) = [7]$ and $E(G) = \{12, 23, 34, 45, 56, 67, 17\}$
2. G such that $V(G) = [7]$ and $E(G) = \{12, 13, 23, 34, 45, 56, 67, 17\}$
3. K_n
4. A subgraph of the triangular lattice in the plane.

Which of the graphs above are perfect?

Problem 6.2. Play the combinatorial secant varieties “game” with the rational normal curve = binomial random variable. The generators of the secant ideal will be minors of Hankel matrices. The appropriate poset is a “zigzag”.