Lecture 4:

Summary:

* $\underbrace{\text{Discrete exponential families}}_{\text{(DEF)}}$ are toric varieties
* Hypothesis testing for DEF using an exact Fisher test is closely related to toric ideals a.k.a Markov basis.
* Conditional indep. statements for discrete models, in several cases these are also DEF. e.g. decomposable models.
* Today: Maximum likelihood degree

Ch5, Ch7. Sullivant, Lectures in Algebraic Statistics.

<u>Def</u>: Let $D = \{X^{(1)}, \dots, X^{(n)}\}$ be data from the same model with parameter space Θ. For discrete data, the likelihood function is

$$L(\theta \mid D) := p_\theta(D).$$

→ What is the prob. of observing D, if θ is the parameter.
→ This is a function of θ, with D fixed.

<u>Def</u>: The maximum likelihood estimate (MLE) $\hat{\theta}$ is the maximizer of the likelihooh function

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta \mid D)$$

<u>Prop</u>: Let $\mathcal{M}_{X \perp\!\!\!\perp Y}$ be the indep. model. $\mathcal{X} = [r_1]$, $\mathcal{Y} = [r_2]$. Let $u \in \mathbb{N}^{r_1 \times r_2}$ be the table of counts, obtained from i.i.d samples

Let

$$\begin{array}{c|c|c} & Y & \\ \hline X & u & \begin{matrix} u_{1+} \\ \vdots \\ u_{r_1+} \end{matrix} \\ \hline & u_{+1} \ \dots \ u_{+r_2} & u_{++} \end{array}$$

be the table of marginals and $n = u_{++}$ the sample size.

$$\hat{p}_{i_1 i_2} = \left(\frac{u_{i_1+}}{n}\right)\left(\frac{u_{+i_2}}{n}\right)$$

When $D = X^{(1)}, \dots, X^{(n)}$ are i.i.d samples, then

$$\mathcal{L}(\theta | D) = \mathcal{L}(\theta / X^{(1)}, \dots, X^{(n)})$$

$$= p_\theta(X^{(1)}, \dots, X^{(n)})$$

$$= \prod_{i=1}^{n} p_\theta(X^{(i)}) = \prod_{j \in [r]} p_\theta(X=j)^{u_j}$$

→ When D is summarized as $u = (u_1, \dots, u_r)$, ⇒ multiply times coeff = $\binom{n}{u}$ → multinomial.

→ log-likelihood. $\ell(\theta | D) = \log \mathcal{L}(\theta / D) = \sum_{j=1}^{r} u_j \log(p_j(\theta))$

One way to find the MLE is to look at the solutions of the score equations $\nabla \ell(\theta | D) = 0$ because max occurs at a critical pt.

$$\left\{ \frac{\partial}{\partial \theta_i} \ell(\theta | D) = 0 \right. , \quad i = 1, \dots, d.$$

$$\Leftrightarrow \left\{ \sum_{j=1}^{r} \frac{u_j}{p_j} \frac{\partial p_j(\theta)}{\partial \theta_i(\theta)} \right. , \quad i = 1, \dots, d.$$

<u>Thm:</u> Let $\mathcal{M}_\theta \subseteq \Delta_{r-1}$ be a statistical model. For generic data, the number of solutions of the score equations is independent of u.
pf. See Sullivant, Algebraic Stat. p. 139.

<u>Rmk:</u> This is a theorem over $\mathbb{C}$, $u \in \mathbb{N}^r$, but $u \in \mathbb{C}^r$ is OK.
Generic means the theorem holds in the complement of a proper subvariety of $\mathbb{C}^r$.

<u>Def:</u> The number of solutions to the score equations for generic u is called the "maximum likelihood degree" of the statistical model.

Examples:

(1) $\mathcal{M}_{X \perp\!\!\!\perp Y}$ has MLD = 1, $\hat{p}_i$ is infact a rational function of u_i.

Having MLD = 1 is the same as saying $\hat{p}_i$ is a rational function of the data.

(2) DEFs do not in general have MLD = 1. Only in some instances, e.g. Decomposable models.

(3) MLEs for DEF exist and can be compute nicely via.
Algorithm: Iterative proportional fitting/scaling.

MLE for log-linear models, A integer matrix $\mathbb{1} \in \text{rowspan}(A)$

<u>Prop:</u> Let $A \in \mathbb{N}^{d \times k}$, $u \in \mathbb{N}^k$ be a vector of positive counts.
The MLE of the frequencies $\hat{u}_i = n\hat{p}_i$ in the log-linear model $\mathcal{M}_A$ is the unique non-negative solution to the simultaneous system of equations

$$A\hat{u} = Au \quad \text{and} \quad \hat{u} \in V(I_A)$$

pf. Use Lagrange multipliers and log-likelihood recall that $p \in \mathcal{M}_A \Leftrightarrow \log(p) = xA$, $x \in \text{rowspan}(A)$.

Iterative proportional fitting. Suppose A has the property that all column sums are a.

<u>INPUT:</u> Matrix $A \in \mathbb{N}^{k \times r}$, a vector $h \in \mathbb{R}^r_{>0}$, counts $u \in \mathbb{N}^r$, tolerance, $\varepsilon > 0$.
<u>OUTPUT:</u> MLE $\hat{p} \in \mathcal{M}_{A,h}$ given u.

<u>Step 1:</u> Initialize a distribution $p \in \mathcal{M}_{A,h}$. e.g. $= \frac{h_i}{\sum h_i}$

<u>Step 2:</u> While $\|Ap - A\frac{u}{n}\| > \varepsilon$ do: For all $i \in [r]$ set $p_i := p_i \cdot \left(\frac{\phi_i^{A,h}(Au/n)}{\phi_i^{A,h}(Ap)} \right)^{1/a}$

<u>Step 3:</u> Output $\hat{p} = p$.