Likelihood Geometry & Intro to exact testing for log-linear models

Algebraic & Geometric Methods in Statistics

Sonja Petrović
Created for Math/Stat 561

Feb 19, 2026.

Related readings

Chapter 7 from our textbook.

Goals

- Understand examples
- Understand counting the number of solutions
- See how it all plays out in the discrete exponentail family case.

Likelihood geometry

- Recap: Likelihood inference

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \{p = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \in \Delta_3 : p_{ij} = \alpha_i \beta_j, (\alpha, \beta) \in \Delta_1 \times \Delta_1\} \text{ and } u = \begin{pmatrix} 19 & 141 \\ 17 & 149 \end{pmatrix}$$

Log-likelihood function: $l(\alpha, \beta \,|\, u) = 160 \log \alpha_1 + 166 \log \alpha_2 + 36 \log \beta_1 + 290 \log \beta_2$

$$= 160 \log \alpha_1 + 166 \log(1 - \alpha_1) + 36 \log \beta_1 + 290 \log(1 - \beta_1)$$

Score equations:

$$\frac{\partial l(\alpha, \beta \,|\, u)}{\partial \alpha_1} = \frac{160}{\alpha_1} - \frac{166}{1 - \alpha_1} = 0$$

$$\frac{\partial l(\alpha, \beta \,|\, u)}{\partial \beta_1} = \frac{36}{\beta_1} - \frac{290}{1 - \beta_1} = 0$$

Figure 1: Example of score equtions

Discrete setup

- Parametric model given by a *rational map* $p : \Theta \to \Delta_{r-1}$
- *iid* samples $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ such that $X^{(i)} \sim p$ for some **unknown** p
- The vector of **counts** $u \in \mathbb{N}^r$, with $u_j = |\{i : X^{(i)} = j\}|$
- *Log-likelihood function* $\ell(\theta|u) = \sum_{j=1}^{r} u_j \log p_j$
- **Score equations** $\sum_{j=1}^{r} \frac{u_j}{p_j} \frac{dp_j}{d\theta_i}$. *One equation for each* θ_i*.*

Discrete setup

- Parametric model given by a *rational map* $p : \Theta \to \Delta_{r-1}$
- *iid* samples $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ such that $X^{(i)} \sim p$ for some **unknown** p
- The vector of **counts** $u \in \mathbb{N}^r$, with $u_j = |\{i : X^{(i)} = j\}|$
- *Log-likelihood function* $\ell(\theta|u) = \sum_{j=1}^{r} u_j \log p_j$
- **Score equations** $\sum_{j=1}^{r} \frac{u_j}{p_j} \frac{dp_j}{d\theta_i}$. *One equation for each* θ_i.

Theorem & Definition

Let $\mathcal{M} \subseteq \Delta_{r-1}$ be a statistical model. For *generic*[a] data, the number of solutions to the score equations is independent of u.
The number of solutions to the score equations for generic u is called the **maximum likelihood degree** (ML degree) of the parametric discrete statistical model $\mathcal{M}$.

[a] 'sufficiently random', outside a variety

Computational algebra is really good at coutning the number of solutions to

Implicit models

Problem

Given vector of counts u, we would like to maximize the log-likelihood function $\ell(\theta|u) = \sum_{j=1}^{r} u_j \log p_j$ over the *intersection* of the interior of the probability simplex Δ_{r-1} and the variety V(polynomials defining the model).

Example

$$M_{X \perp\!\!\!\perp Y} = \{ \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \in \Delta_3 : p_{11}p_{22} - p_{12}p_{21} = 0 \} \text{ and } u = \begin{bmatrix} 19 & 141 \\ 17 & 149 \end{bmatrix}.$$

- Maximize $\ell(p|u) = 19 \log p_{11} + 141 \log p_{12} + 17 \log p_{21} + 149 \log p_{22}$ over $M_{X \perp\!\!\!\perp Y}$.
- The polynomial **constraints** are $p_{11} + p_{12} + p_{21} + p_{22} = 1$ and $p_{11}p_{22} - p_{12}p_{21} = 0$.

$\rightarrow$ Go to lecture11-interlude-LangrangeMultipliers.pdf

Exponential families have concave likelihood functions

Proposition

Let $\mathcal{M}$ be an *exponential family* with minimal sufficient statistics $T(x)$ and natural parameter η. ($f_\eta(x) = h(x)e^{\eta^t T(x) - A(\eta)}$.) Then the likelihood function is strictly concave.

- The MLE, if it exists, is the solution to $T(x) = \mathbb{E}_\eta[T(X)]$.
  - x denotes the data vector.

- *iid* samples $\implies$ sufficient statistic of the sample is $T_n(X^{(1)}, \dots, X^{(n)}) = \sum_{i=1}^{n} T(X^{(i)})$.

Theorem (Prop 7.3.7)

Exponential family $p_\theta(x) = h(x)\exp(\langle\theta, T(x)\rangle - A(\theta))$ *with* sufficient statistics $T(x)$, log-partition function $A(\theta) = \log \int_{\mathcal{X}} h(x) exp(\langle\theta, T(x)\rangle)$ *Then*

$$\frac{\partial}{\partial\theta_i} A(\theta) = \mathbb{E}_\theta[T_i(X)] \quad and \quad \frac{\partial^2}{\partial\theta_i\theta_j} A(\theta) = \mathrm{Cov}_\theta[T_i(X), T_j(X)].$$

Example: ML degree of (rescaled) binomial is 3

$$p(\theta) = (s, s\theta, s\theta^2, s\theta^3) \subset \Delta_3 \subset \mathbb{R}^4.$$

where $s = \frac{1}{1+\theta+\theta^2+\theta^3}$. Sample size $n = u_0 + u_1 + u_2 + u_3$. We have

$$\begin{aligned} L(\theta|u) &= s^{u_0}(s\theta)^{u_1}(s\theta^2)^{u_2}(s\theta^3)^{u_3} \\ &= s^{u_0+u_1+u_2+u_3}\theta^{u_1+2u_2+3u_3} \end{aligned}$$

$$\ell(\theta|u) = n \log s + (u_1 + 2u_2 + 3u_3) \log \theta$$

The score equation is:

$$0 = \frac{\partial \ell}{\partial \theta} = -ns(1 + 2\theta + 3\theta^2) + (u_1 + 2u_2 + 3u_3)\frac{1}{\theta}$$

Thus $3n\theta^3 + 2n\theta^2 + n\theta - (u_1 + 2u_2 + 3u_3)s^{-1} = 0$ and we arrive at

$$3(n - u_3)\theta^3 + 2(n - u_2)\theta^2 + (n - u_1)\theta - (u_1 + 2u_2 + 3u_3) = 0$$

ML for discrete expo fam.

Theorem (Prop 7.3.7)

Exponential family $p_\theta(x) = h(x)\exp(\langle\theta, T(x)\rangle - A(\theta))$ *with* sufficient statistics $T(x)$, log-partition function $A(\theta) = \log\int_{\mathcal{X}} h(x) exp(\langle\theta, T(x)\rangle)$ *Then*

$$\frac{\partial}{\partial\theta_i}A(\theta) = \mathbb{E}_\theta[T_i(X)] \quad \textit{and} \quad \frac{\partial^2}{\partial\theta_i\theta_j}A(\theta) = \mathrm{Cov}_\theta[T_i(X), T_j(X)].$$

Corollary (Cor 7.3.8)

The likelihood function for an exponential family is strictly concave. The MLE (if it exists) is the unique solution to the equation

$$\mathbb{E}_\theta[T(X)] = T(x)$$

where x denotes the data vector.

Figure 4: Source: Carlos Amendola

Corollary [Birch's theorem]

Let $A \subseteq \mathbb{Z}^{k \times r}$ such that $1 \in rowspan(A)$. Let u be a vector of counts from *iid* samples. Then the MLE of the log-linear model is the unique solution, if it exists, to

$$Au = nAp \text{ and } p \in \mathcal{M}_A.$$

- Inspires algorithms for computing MLE:
  - Iterative proportional fitting. Stephen Fienberg, AMS 1970.
- R can do this - it's super fast
  - some resources at end of these slides
  - IPF is usually embedded inside other functions

```
fm <- loglin(HairEyeColor, list(c(1, 2), c(1, 3), c(2, 3)))
```

```
5 iterations: deviation 0.04093795
```

```
## fm
```

The following problem will appear on HW 3

- exercise 7.2. in the book

Let $\mathcal{M}$ be the model of binomial random variables $Bin(2, \theta)$:

$$\mathcal{M} = \{(1-\theta)^2, 2\theta(1-\theta), \theta^2) \in \Delta_2 : \theta \in (0,1)\}.$$

- What is the ML degree of $\mathcal{M}$?
- Compute the MLE $\hat{\theta}$ for the two data points $u = (8, 6, 5)$ and $v = (4, 20, 8)$. Interpret your results

Interlude: log-linear models (in 2023, this was a long class discussion)

Did anyone try and succeed to write out what it means that "log(p) is in the rowspan(A)" for the example of the independence model?

Interlude: log-linear models (in 2023, this was a long class discussion)

Observation

Let $p \in \mathcal{M}_{X \perp\!\!\!\perp Y}$. If p has all positive entries ($p \in \mathrm{int}(\Delta_{\mathcal{R}-1})$) then

$$
\begin{aligned}
\log p &= (\log p_{1+}p_{+1}, \log p_{1+}p_{+2}, \log p_{2+}p_{+1}, \log p_{2+}p_{+2}) \\
&= (\log p_{1+} + \log p_{+1}, \log p_{1+} + \log p_{+2}, \log p_{2+} + \log p_{+1}, \log p_{2+} + \log p_{+2}) \\
&= \log p_{1+}(1,1,0,0) + \log p_{2+}(0,0,1,1) + \log p_{+1}(1,0,1,0) \\
&+ \log p_{+2}(0,1,0,1).
\end{aligned}
$$

Thus $\log p \in \mathcal{M}_A$, where $A \in \mathbb{Z}^{4\times 4}$ is the matrix

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

- Answer by Miles:

In general (from slide 13 or lec. 7):

$p_\theta = \frac{1}{Z(\theta)} h \prod_j \theta^{a_j}$ where a_j is the j^{th} row of $A \in Z^{k \times r}$.

If $p_\theta \in \text{int}(\triangle_{\mathcal{R}-1})$ then $(1, \ldots, 1) = \mathbf{1} \in \text{rowspan}(\mathbf{A})$ i.e. $\mathbf{1} = \mathbf{cA}$ for some vector $c \in \mathbb{Z}^r$

Assume $h = \mathbf{1}$. Then $\log p_\theta = \log(h) - \log(Z(\theta))\mathbf{1} + \sum_{\mathbf{j}} \mathbf{a_j} \log \theta$
$= \mathbf{0} - \log(\mathbf{Z}(\theta))\mathbf{cA} + \log \theta \mathbf{A} \qquad = (-\log(Z(\theta))c + \log \theta)\, A$

Here $-\log(Z(\theta))c + \log \theta$ is just a vector, in $\mathbb{R}^r$ so this means $\log p_\theta \in \text{rowspan}(A)$

Exact testing!

this is our next topic!!

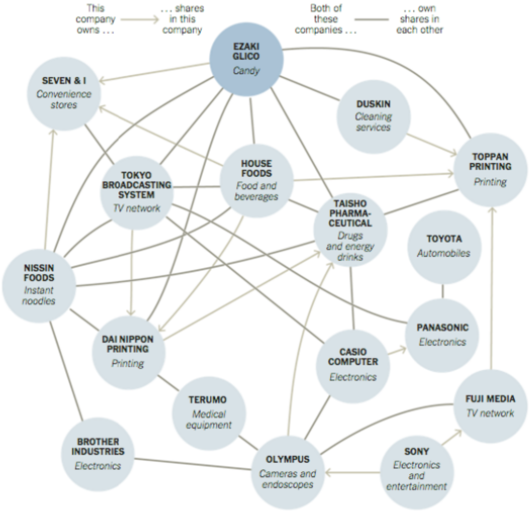
- Last slide from likelihood geometry said: “IPF is usually embedded inside other functions”
- . . . which begs the question: What other questions might we have??

The following few slides are a **preview** of what kinds of questions we can answer with our next topic.

Is the given set of shareowning relations expected or not?

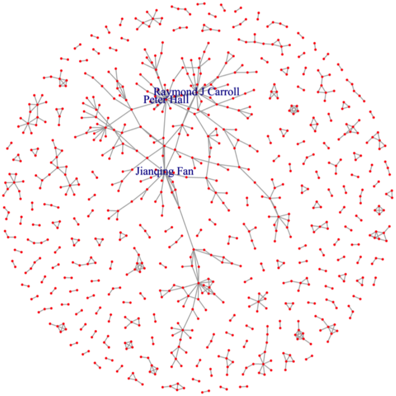
Inside Japan Inc.

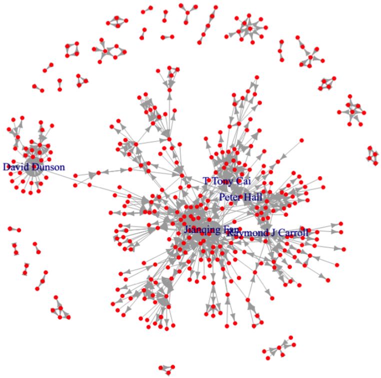
Many companies in Japan own shares in each other to create relationships that can protect them from outside intereference. Here are some companies that have disclosed their connections, beginning with Ezaki Glico, a candy maker that has struggled to post steady returns even as it has resisted other shareholders' demands for change.



Sources: Company reports

Are degrees a good summary of a network?





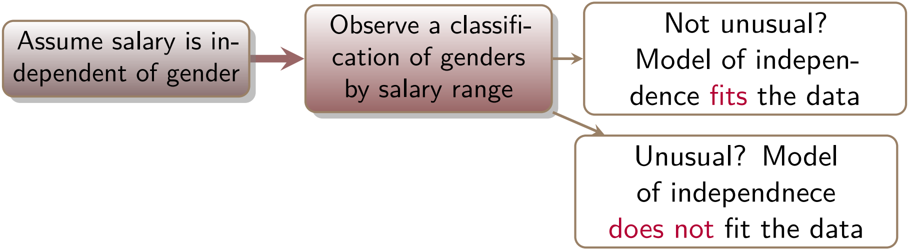
At the heart of statistical reasoning

- Given: data, find out if it is usual/expected? surprising/outlier? quantify??

Formal reasoning with data: independence example

- **Modeling**: Construct a statistical model for independence.
- **Question**: Does the model fit the observed set of gender vs. salary ranges?

  (Can it adequately explain how the salary data was generated?)
- **Process**:



Conclusions

Main take-aways about *likelihood geometry*

- Numerical algorithms for computing MLE, for example the EM algoritm implemented widely, are usually some form of hill-climbing. They have no way of telling you whether you are at a global or local optimum.
- Likelihood function in exponential families is strictly concave
  - However there can be local optima on the boundary of the model
- When you compute estimates numerically, it is a good idea to understand how many critical points there are
  - You can set up the system of score equations
  - You can count the number of (complex) solutions to those equations
    - This quantity, called the ML degree in algebraic statisitcs, is one measure of complexity of estimation.
- ML degree is one *if and only if* the MLE formula is a rational function of the data.
  - Birch's theorem.

Additional material

- Here is a vignette about how IPF algorithm works in `R`.
- In `python`, I have not used this, but found this link which appears to be useful: IPF in python

License

Parts of this presentation are from Kaie Kubjas' course lectures, used with permission; and Carlos Amendola's lecture in Bernd Sturmfel's short course on Algebraic Statistics in Berlin, fall 2022.

This document is created for Math/Stat 561, Spring 2023.