week 12 day 1

Graphical models: MLE

Algebraic & Geometric Methods in Statistics

Sonja Petrović
Created for Math/Stat 561

April 3, 2023.

Agenda

- Computing the MLE of an example graph $\mapsto$ homework 5

Figure 3.2.3: Directed graphs representing (a) $X_1 \perp\!\!\!\perp X_3 \mid X_2$ and (b) $X_1 \perp\!\!\!\perp X_2$.

Figure 1: Source: Oberwolfach lectures

- Here is an incredible online resource: Maathuis, Drton, Lauritzen & Wainwright's Handbook of graphical models

Graphical models part 3: how to compute MLEs

- Computing the MLE of an example graph $\mapsto$ homework 5

pages 2-13

Example: Risk Factors for Coronary Heart Disease

Data collected from a sample of 1841 workers employed in the Czech automotive industry.

- S: smoked
- B: systolic blood pressure was less than 140 mm
- H: family history of coronary heart disease
- L: ratio of beta to alpha lipoproteins less than 3

Example: Risk Factors for Coronary Heart Disease

Data collected from a sample of 1841 workers employed in the Czech automotive industry.

- S: smoked
- B: systolic blood pressure was less than 140 mm
- H: family history of coronary heart disease
- L: ratio of beta to alpha lipoproteins less than 3

Random vector $X = (S, B, H, L)$ with each risk factor a *binary* variable, so X has a state space of cardinality 16:

$$p_{ijkl} = Prob(S = i, B = j, H = k, L = l), \qquad i, j, k, l \in 0, 1.$$

Data

H	L	B	S: no	S: yes
neg	< 3	< 140	297	275
		≥ 140	231	121
	≥ 3	< 140	150	191
		≥ 140	155	161
pos	< 3	< 140	36	37
		≥ 140	34	30
	≥ 3	< 140	32	36
		≥ 140	26	29

Data

H	L	B	S: no	S: yes
neg	< 3	< 140	297	275
		≥ 140	231	121
	≥ 3	< 140	150	191
		≥ 140	155	161
pos	< 3	< 140	36	37
		≥ 140	34	30
	≥ 3	< 140	32	36
		≥ 140	26	29

$$(u_{ijkl} : i, j, k, l \in 0, 1) = (u_{0000}, u_{0001}, ..., u_{1111}) = (297, 275, ..., 29)$$

Question

$(u_{ijkl} : i, j, k, l \in 0, 1) = (u_{0000}, u_{0001}, ..., u_{1111}) = (297, 275, ..., 29)$

$p_{ijkl} = Prob(S = i, B = j, H = k, L = l), \qquad i, j, k, l \in 0, 1.$

Given the observed table, what is the probability distribution $\hat{p} = (\hat{p}_{ijkl})$ that "best" explains the data ?

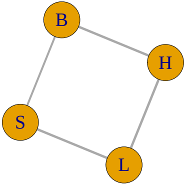
Remember:

- S: smoked
- B: systolic blood pressure was less than 140 mm
- H: family history of coronary heart disease
- L: ratio of beta to alpha lipoproteins less than 3

Maximum likelihood estimation

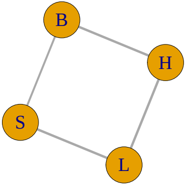
- Pre-specified probability model $\mathcal{M}$ — a subset of all possible probability distributions.
- Choose $\hat{p}$ from $\mathcal{M}$.

Example model [Binary 4-cycle]



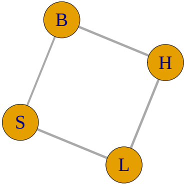
- Model parameters are:

Example model [Binary 4-cycle]



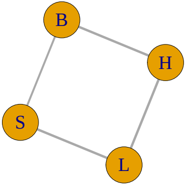
- Model parameters are: $a_{ij}, b_{jk}, c_{kl}, d_{il}$, for $i, j, k, l \in \{0, 1\}$
- $p_{ijkl} = a_{ij} b_{jk} c_{kl} d_{il}$

Example model [Binary 4-cycle]



- Model parameters are: $a_{ij}, b_{jk}, c_{kl}, d_{il}$, for $i, j, k, l \in \{0, 1\}$
- $p_{ijkl} = a_{ij} b_{jk} c_{kl} d_{il}$
- $\mathcal{M}$ is the set of all probability distributions that can be parametrized in this way.
- Distributions in M have the following properties:

Example model [Binary 4-cycle]



- Model parameters are: $a_{ij}, b_{jk}, c_{kl}, d_{il}$, for $i, j, k, l \in \{0, 1\}$
- $p_{ijkl} = a_{ij} b_{jk} c_{kl} d_{il}$
- $\mathcal{M}$ is the set of all probability distributions that can be parametrized in this way.
- Distributions in M have the following properties:
  - S and H are independent given B and L.
  - B and L are independent given S and H.

Maximum likelihood estimation

- Likelihood function

$$\ell_u(p) = \prod_{i,j,k,l} p_{ijkl}^{u_{ijkl}}.$$

- Look for the maximizer $\hat{p} = (\hat{p}_{ijkl})$:

$$\text{maximize } \ell_u(p) = p_{0000}^{u_{0000}} p_{0001}^{u_{0001}} \cdots p_{1111}^{u_{1111}} \text{ subject to } p = (p_{ijkl}) \in \mathcal{M}$$

- The optimal solution $\hat{p}$ is the MLE, the *maximum likelihood estimate* (of the data u for the model $\mathcal{M}$).

Homework 5 problem:

Compute this value $\hat{p}$ explicitly. Using software, by hand, whatever you like!

MLE computation option: score equations

- most straightforward given the one example
- write log-likelihood
- take partial derivatives
- solve (probably numerically using some software of your choice? submit your code!)

Computing the MLE of a Parametrized Statistical Model

- Model parametrized by $\psi : \mathcal{U} \subset \mathbb{R}^d \longrightarrow \mathcal{M} \subset \mathbb{R}^n$:

$$\theta = (\theta_1, \ldots, \theta_d) \mapsto (f_1(\theta), f_2(\theta), \ldots, f_n(\theta))$$

- Observed data $u = (u_1, u_2, \ldots, u_n)$ with sample size $N = \sum u_i$.
- maximize $\ell_u(\theta) = f_1^{u_1} f_2^{u_2} \cdots f_n^{u_n}$ subject to $f_1 + f_2 + \cdots + f_n = 1$.
- maximize $\log \ell_u(\theta) = u_1 \log f_1 + u_2 \log f_2 + \cdots + u_n \log f_n$ subject to $f_1 + f_2 + \cdots + f_n = 1$.

The Likelihood Equations

- maximize $\log \ell_u(\theta) = u_1 \log f_1 + u_2 \log f_2 + \cdots + u_n \log f_n$ subject to $f_1 + f_2 + \cdots + f_n = 1$.
- Compute the critical points of $\log \ell_u(\theta)$. That is, solve the *likelihood equations* (where μ is the Lagrange multiplier):

$$\begin{aligned}
\frac{1}{\ell_u(\theta)} \cdot \frac{\partial \ell_u(\theta)}{\partial \theta_1} &= \mu \frac{\partial}{\partial \theta_1}(f_1 + \cdots + f_n - 1) \\
\frac{1}{\ell_u(\theta)} \cdot \frac{\partial \ell_u(\theta)}{\partial \theta_2} &= \mu \frac{\partial}{\partial \theta_2}(f_1 + \cdots + f_n - 1) \\
\vdots \quad &= \quad \vdots \\
\frac{1}{\ell_u(\theta)} \cdot \frac{\partial \ell_u(\theta)}{\partial \theta_d} &= \mu \frac{\partial}{\partial \theta_d}(f_1 + \cdots + f_n - 1) \\
1 &= f_1 + f_2 + \cdots + f_n
\end{aligned}$$

- The best critical point $\hat{\theta}$ is the MLE.

Some of the theory behind MLE computation

- In general for many models there is no analytic formula for the MLE.
- Finding a local maximum of the likelihood function by numerical hill climbing-type methods $\leftarrow$ most popular in practice!
- Typical problems: not finding global maximum, slow convergence...

Definition (informal)

The maximum likelihood degree (ML degree) of an algebraic statistical model is the number of complex critical points of the likelihood equations for generic data u.

- ML degree is a measure of complexity for maximum likelihood estimation problem for a model.
- ML degree is one $\iff$ the MLE is a rational function of the data.
  - ML Degree of Binary Four Cycle: **13**.

Epilogue

- What other options do we have for computing the MLE in this example?

MLE computation option: Lagrange multipliers

Recall that the method of Lagrange multipliers is used to solve the following constrained optimization problem:

$$\max f(x) \text{ subject to } g_i(x) = 0, i = 1, \ldots, k.$$

MLE computation option: Lagrange multipliers

Recall that the method of Lagrange multipliers is used to solve the following constrained optimization problem:

$$\max f(x) \text{ subject to } g_i(x) = 0, i = 1, \ldots, k.$$

- In the example: $f(x) = \ell_u(p)$, $g_i(x) =$ the polynomials that define the CI ideal of the graphical model:

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

Recall that the method of Lagrange multipliers is used to solve the following constrained optimization problem:

$$\max f(x) \text{ subject to } g_i(x) = 0, i = 1, \ldots, k.$$

- The **Lagrangian** of this optimization problem is

$$L(x, \lambda) = f(x) - \sum_{i=1}^{k} \lambda_i g_i(x).$$

Recall that the method of Lagrange multipliers is used to solve the following constrained optimization problem:

$$\max f(x) \text{ subject to } g_i(x) = 0, i = 1, \ldots, k.$$

- The **Lagrangian** of this optimization problem is

$$L(x, \lambda) = f(x) - \sum_{i=1}^{k} \lambda_i g_i(x).$$

- In the example:

$$L(p, \lambda) = \sum_{i,j,k,l} u_{ijkl} \log p_{ijkl} - \lambda_0 \left(\sum_{ijkl} p_{ijkl} - 1 \right) - \sum_{m=1}^{k} \lambda_m g_m(p).$$

(k = number of binomials)

- The constrained critical points of f are among the unconstrained critical points of L. Hence one has to solve:

$g_1 = 0 \ldots g_k = 0$ for those binomials g_i on the previous slide!

and

$$\frac{\partial f}{\partial x_1} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial x_1} = 0, \ldots, \frac{\partial f}{\partial x_m} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial x_r} = 0$$

MLE computation option: Discrete exponential families

Corollary [from an old lecture!] — Birch's Theorem

$A \subset \mathbb{Z}^{k\times r}$ such that $1 \in rowspan(A)$. $h \in \mathbb{R}^r_{>0}$ and u vector of counts from n iid samples.

Then the MLE of the joint probability vector p in the log-linear model $\mathcal{M}_{A,h}$ given the data u is the unique – if it exists – solution of the equations

$$Au = nAp \text{ and } p \in \mathcal{M}_{A,h}.$$

We can use Birch's theorem with numerical algorithms!

ML geomery group at 2016 MRC

Carlos Amendola, Courtney Gibbons, Evan Nash, Nathan Bliss, Martin Helmer, Jose Rodriguez, Isaac Burke, Serkan Hosten, Daniel Smolkin. arXiv:1703.02251

- Jump to example slides

The usual. . . license

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

Sources: textbook, Carlos Enrique Améndola Cerón's slides from a M2 workshop in GA Tech 2017.