The Maximum Likelihood Degree

Carlos Enrique Améndola Cerón

Technische Universität Berlin

Macaulay2 Tutorials
July 2017, Georgia Tech

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

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:

$$p_{ijk\ell} = \mathrm{prob}(S = i, B = j, H = k, L = \ell) \quad i, j, k, \ell \in \{0, 1\}$$

Risk Factors for Coronary Heart Disease

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

Risk Factors for Coronary Heart Disease

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:

$$(u_{ijk\ell} \;:\; i, j, k, \ell \in \{0, 1\}) = (u_{0000}, u_{0001}, \ldots, u_{1111}) = (297, 275, \ldots, 29)$$

Risk Factors for Coronary Heart Disease

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:
$(u_{ijk\ell} \ : \ i, j, k, \ell \in \{0, 1\}) = (u_{0000}, u_{0001}, \ldots, u_{1111}) = (297, 275, \ldots, 29)$

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

Maximum Likelihood Estimation

Pre-specified probability model $\mathcal{M}$ (a subset of all possible probability distributions).

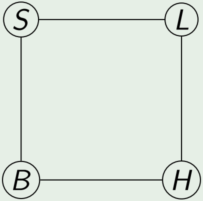
Maximum Likelihood Estimation

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

Maximum Likelihood Estimation

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

Example (Binary Four-Cycle)

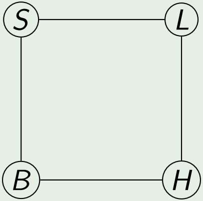


- $a_{ij}, b_{jk}, c_{k\ell}, d_{i\ell}$ parameters for $i, j, k, \ell \in \{0, 1\}$ and let $p_{ijk\ell} = a_{ij} b_{jk} c_{k\ell} d_{i\ell}$
- $\mathcal{M}$ is the image of this parametrization

Maximum Likelihood Estimation

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

Example (Binary Four-Cycle)



- $a_{ij}, b_{jk}, c_{k\ell}, d_{i\ell}$ parameters for $i, j, k, \ell \in \{0, 1\}$ and let $p_{ijk\ell} = a_{ij} b_{jk} c_{k\ell} d_{i\ell}$
- $\mathcal{M}$ is the image of this parametrization
- Distributions in $\mathcal{M}$ have the property that S and H are independent given B and L. Also, B and L are independent given S and H.

Maximum Likelihood Estimation

- Likelihood function

$$\ell_u(p) = \prod_{i,j,k,\ell} p_{ijk\ell}^{u_{ijk\ell}}$$

Maximum Likelihood Estimation

- Likelihood function

$$\ell_u(p) = \prod_{i,j,k,\ell} p_{ijk\ell}^{u_{ijk\ell}}$$

Look for the maximizer $\hat{p} = (\hat{p}_{ijk\ell})$!

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

Maximum Likelihood Estimation

- Likelihood function

$$\ell_u(p) = \prod_{i,j,k,\ell} p_{ijk\ell}^{u_{ijk\ell}}$$

Look for the maximizer $\hat{p} = (\hat{p}_{ijk\ell})$!

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

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

Maximum Likelihood Estimation

- Likelihood function

$$\ell_u(p) = \prod_{i,j,k,\ell} p_{ijk\ell}^{u_{ijk\ell}}$$

Look for the maximizer $\hat{p} = (\hat{p}_{ijk\ell})$!

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

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

Example (computed with M2)

$$\hat{p} = (\, 0.15293342, 0.089760679, 0.021266977, 0.015778191,$$
$$0.12976986, 0.076165372, 0.020853199, 0.015471205,$$
$$0.13533793, 0.11789409, 0.018820142, 0.0207235,$$
$$0.083859917, 0.073051125, 0.01347576, 0.014838619\,).$$

Computing the Maximum Likelihood Estimate

- In general for many models there is no analytic formula for the MLE.

Computing the Maximum Likelihood Estimate

- 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

Computing the Maximum Likelihood Estimate

- 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
- Typical problems: not finding global maximum, slow convergence...

Algebraic Statistics Mantra

Many classes of parametrized probability models are algebraic varieties.

Algebraic Statistics Mantra

Many classes of parametrized probability models are algebraic varieties.

Example

$\psi : (\mathbb{C}^*)^{16} \longrightarrow (\mathbb{C}^*)^{16}$ given by

$$(a_{00}, a_{01}, a_{10}, a_{11}, b_{00}, \ldots, c_{00}, \ldots, d_{00}, \ldots) \mapsto (p_{0000}, p_{0001}, \ldots, p_{1111})$$

where $p_{ijk\ell} = a_{ij} b_{jk} c_{k\ell} d_{i\ell}$ for $i, j, k, \ell \in \{0, 1\}$.

Algebraic Statistics Mantra

Many classes of parametrized probability models are algebraic varieties.

Example

$\psi : (\mathbb{C}^*)^{16} \longrightarrow (\mathbb{C}^*)^{16}$ given by

$$(a_{00}, a_{01}, a_{10}, a_{11}, b_{00}, \ldots, c_{00}, \ldots, d_{00}, \ldots) \mapsto (p_{0000}, p_{0001}, \ldots, p_{1111})$$

where $p_{ijk\ell} = a_{ij} b_{jk} c_{k\ell} d_{i\ell}$ for $i, j, k, \ell \in \{0, 1\}$.

- $V = \overline{\psi((\mathbb{C}^*)^{16})} \subset \mathbb{C}^{16} \subset \mathbb{P}^{15}$ is a projective (toric) variety.

Algebraic Statistics Mantra

Many classes of parametrized probability models are algebraic varieties.

Example

$\psi : (\mathbb{C}^*)^{16} \longrightarrow (\mathbb{C}^*)^{16}$ given by

$$(a_{00}, a_{01}, a_{10}, a_{11}, b_{00}, \ldots, c_{00}, \ldots, d_{00}, \ldots) \mapsto (p_{0000}, p_{0001}, \ldots, p_{1111})$$

where $p_{ijk\ell} = a_{ij} b_{jk} c_{k\ell} d_{i\ell}$ for $i, j, k, \ell \in \{0, 1\}$.

- $V = \overline{\psi((\mathbb{C}^*)^{16})} \subset \mathbb{C}^{16} \subset \mathbb{P}^{15}$ is a projective (toric) variety.
- $\overline{\mathcal{M}} = V$.

Probability Models as Algebraic Varieties

Example (Equations for the Binary Four-Cycle)

The projective variety V corresponding to the binary four-cycle defined by

$$\langle\, p_{1011}p_{1110} - p_{1010}p_{1111},\; p_{0111}p_{1101} - p_{0101}p_{1111},\; p_{1001}p_{1100} - p_{1000}p_{1101},\; p_{0110}p_{1100} - p_{0100}p_{1110},$$
$$p_{0011}p_{1001} - p_{0001}p_{1011},\; p_{0011}p_{0110} - p_{0010}p_{0111},\; p_{0001}p_{0100} - p_{0000}p_{0101},\; p_{0010}p_{1000} - p_{0000}p_{1010},$$
$$p_{0100}p_{0111}p_{1001}p_{1010} - p_{0101}p_{0110}p_{1000}p_{1011},\; p_{0010}p_{0101}p_{1011}p_{1100} - p_{0011}p_{0100}p_{1010}p_{1101},$$
$$p_{0001}p_{0110}p_{1010}p_{1101} - p_{0010}p_{0101}p_{1001}p_{1110},\; p_{0001}p_{0111}p_{1010}p_{1100} - p_{0011}p_{0101}p_{1000}p_{1110},$$
$$p_{0000}p_{0011}p_{1101}p_{1110} - p_{0001}p_{0010}p_{1100}p_{1111},\; p_{0000}p_{0111}p_{1001}p_{1110} - p_{0001}p_{0110}p_{1000}p_{1111},$$
$$p_{0000}p_{0111}p_{1011}p_{1100} - p_{0011}p_{0100}p_{1000}p_{1111},\; p_{0000}p_{0110}p_{1011}p_{1101} - p_{0010}p_{0100}p_{1001}p_{1111}\,\rangle.$$

Computing the MLE of a Parametrized Statistical Model

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

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

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 $N = \sum u_i$.

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 $f_1 + f_2 + \cdots + f_n = 1$.

Computing the MLE of a Parametrized Statistical Model

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

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

- Observed data $u = (u_1, u_2, \dots, u_n)$ 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 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.

Maximum Likelihood Degree

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.

Maximum Likelihood Degree

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.

Maximum Likelihood Degree

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.

Maximum Likelihood Degree

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.

Example (ML Degree of Binary Four Cycle)

The ML degree of the binary four cycle is 13.

Example (Twisted Cubic Model)

$$\phi(s,t) = (s, st, st^2, st^3) \subset \Delta_4 \subset \mathbb{R}^4.$$

Example (Twisted Cubic Model)

$$\phi(s,t) = (s, st, st^2, st^3) \subset \Delta_4 \subset \mathbb{R}^4.$$

The likelihood function is

$$\begin{aligned}\ell_u(s,t) &= s^{u_0}(st)^{u_1}(st^2)^{u_2}(st^3)^{u_3} \\ &= s^{u_0+u_1+u_2+u_3}t^{u_1+2u_2+3u_3}\end{aligned}$$

$$\log \ell_u(s,t) = (u_0+u_1+u_2+u_3)\log s + (u_1+2u_2+3u_3)\log t$$

Example (Twisted Cubic Model)

$$\phi(s,t) = (s, st, st^2, st^3) \subset \Delta_4 \subset \mathbb{R}^4.$$

The likelihood function is

$$\begin{aligned}\ell_u(s,t) &= s^{u_0}(st)^{u_1}(st^2)^{u_2}(st^3)^{u_3} \\ &= s^{u_0+u_1+u_2+u_3} t^{u_1+2u_2+3u_3}\end{aligned}$$

$$\log \ell_u(s,t) = (u_0+u_1+u_2+u_3)\log s + (u_1+2u_2+3u_3)\log t$$

The likelihood equations are:

$$\begin{aligned}\frac{u_0+u_1+u_2+u_3}{s} &= \mu(1+t+t^2+t^3) \\ \frac{u_1+2u_2+3u_3}{t} &= \mu(s+2st+3st^2) \\ s+st+st^2+st^3 &= 1\end{aligned}$$

Example (Twisted Cubic Model)

$$\phi(s,t) = (s, st, st^2, st^3) \subset \Delta_4 \subset \mathbb{R}^4.$$

The likelihood function is

$$\begin{aligned}\ell_u(s,t) &= s^{u_0}(st)^{u_1}(st^2)^{u_2}(st^3)^{u_3} \\ &= s^{u_0+u_1+u_2+u_3} t^{u_1+2u_2+3u_3}\end{aligned}$$

$$\log \ell_u(s,t) = (u_0 + u_1 + u_2 + u_3)\log s + (u_1 + 2u_2 + 3u_3)\log t$$

The likelihood equations are:

$$\begin{aligned}\frac{u_0 + u_1 + u_2 + u_3}{s} &= \mu(1 + t + t^2 + t^3) \\ \frac{u_1 + 2u_2 + 3u_3}{t} &= \mu(s + 2st + 3st^2) \\ s + st + st^2 + st^3 &= 1\end{aligned}$$

ML degree is

Example (Twisted Cubic Model)

$$\phi(s,t) = (s, st, st^2, st^3) \subset \Delta_4 \subset \mathbb{R}^4.$$

The likelihood function is

$$\begin{aligned} \ell_u(s,t) &= s^{u_0}(st)^{u_1}(st^2)^{u_2}(st^3)^{u_3} \\ &= s^{u_0+u_1+u_2+u_3} t^{u_1+2u_2+3u_3} \end{aligned}$$

$$\log \ell_u(s,t) = (u_0+u_1+u_2+u_3)\log s + (u_1+2u_2+3u_3)\log t$$

The likelihood equations are:

$$\begin{aligned} \frac{u_0+u_1+u_2+u_3}{s} &= \mu(1+t+t^2+t^3) \\ \frac{u_1+2u_2+3u_3}{t} &= \mu(s+2st+3st^2) \\ s+st+st^2+st^3 &= 1 \end{aligned}$$

ML degree is 3.

Maximum Likelihood Degree

Definition (Precise)

Let $V \subset \mathbb{P}^{n-1}$ be a projective variety over $\mathbb{C}$, and let

$$\ell_u = \frac{p_1^{u_1} p_2^{u_2} \cdots p_n^{u_n}}{(p_1 + \cdots + p_n)^{(u_1 + \cdots + u_n)}}.$$

The ML degree of V is the number of complex critical points of ℓ_u on $V_{reg} \setminus \mathcal{H}$ for generic data $u = (u_1, \ldots, u_n)$ where

$$\mathcal{H} = \{p \ : \ p_1 \cdots p_n (p_1 + \cdots + p_n) = 0\}.$$

ML Degree: some History

- Catanese-Hoșten-Khetan-Sturmfels [06]: introduced and proved ML degree well-defined
  - if $f_1(\theta), \dots, f_n(\theta)$ are polynomials with generic coefficients, then ML degree is the top Chern class of $\Omega^1_V(\log D)$.
  - under some restricted assumptions ML degree of V is $\pm\chi_{\mathrm{top}}(\mathbb{P}^d \setminus D)$.

ML Degree: some History

- Catanese-Hoșten-Khetan-Sturmfels [06]: introduced and proved ML degree well-defined
  - if $f_1(\theta), \ldots, f_n(\theta)$ are polynomials with generic coefficients, then ML degree is the top Chern class of $\Omega^1_V(\log D)$.
  - under some restricted assumptions ML degree of V is $\pm\chi_{\mathrm{top}}(\mathbb{P}^d \setminus D)$.
- Hoșten-Khetan-Sturmfels [05]: symbolic algorithms to compute ML degree

ML Degree: some History

- Catanese-Hoșten-Khetan-Sturmfels [06]: introduced and proved ML degree well-defined
  - if $f_1(\theta), \ldots, f_n(\theta)$ are polynomials with generic coefficients, then ML degree is the top Chern class of $\Omega^1_V(\log D)$.
  - under some restricted assumptions ML degree of V is $\pm\chi_{\mathrm{top}}(\mathbb{P}^d \setminus D)$.
- Hoșten-Khetan-Sturmfels [05]: symbolic algorithms to compute ML degree
- Hauenstein-Rodriguez-Sturmfels [12]: computed ML degree of various determinantal varieties using NAG

ML Degree: some History

- Catanese-Hoşten-Khetan-Sturmfels [06]: introduced and proved ML degree well-defined
  - if $f_1(\theta), \ldots, f_n(\theta)$ are polynomials with generic coefficients, then ML degree is the top Chern class of $\Omega^1_V(\log D)$.
  - under some restricted assumptions ML degree of V is $\pm\chi_{\mathrm{top}}(\mathbb{P}^d \setminus D)$.
- Hoşten-Khetan-Sturmfels [05]: symbolic algorithms to compute ML degree
- Hauenstein-Rodriguez-Sturmfels [12]: computed ML degree of various determinantal varieties using NAG
- Huh [13]: the ML degree of a smooth very affine variety is $\pm\chi_{\mathrm{top}}(\cdot)$.
- Huh [13]: characterized varieties of ML degree one

Scaled Toric Models

- Integer matrix A of size $(d-1) \times n$

Scaled Toric Models

- Integer matrix A of size $(d-1) \times n$
- Map $\psi_A : (\mathbb{C}^*)^d \longrightarrow (\mathbb{C}^*)^n$ where

$$\psi_A(s, \theta_1, \ldots, \theta_{d-1}) = (s\theta^{a_1},\, s\theta^{a_2},\, \ldots, s\theta^{a_n}).$$

Scaled Toric Models

- Integer matrix A of size $(d-1) \times n$
- Map $\psi_A : (\mathbb{C}^*)^d \longrightarrow (\mathbb{C}^*)^n$ where

$$\psi_A(s, \theta_1, \ldots, \theta_{d-1}) = (s\theta^{a_1}, s\theta^{a_2}, \ldots, s\theta^{a_n}).$$

- *Toric Variety* V_A defined by image of ψ_A.

Scaled Toric Models

- Integer matrix A of size $(d-1) \times n$
- Map $\psi_A \ : \ (\mathbb{C}^*)^d \longrightarrow (\mathbb{C}^*)^n$ where

$$\psi_A(s, \theta_1, \ldots, \theta_{d-1}) = (s\theta^{a_1},\, s\theta^{a_2},\, \ldots, s\theta^{a_n}).$$

- *Toric Variety* V_A defined by image of ψ_A.
- Now, scaling vector $c \in (\mathbb{C}^*)^n$:

$$\psi_A^c(s, \theta_1, \theta_2, \ldots, \theta_{d-1}) \ = \ (c_1 s\theta^{a_1}, c_2 s\theta^{a_2}, \ldots, c_n s\theta^{a_n})$$

Scaled Toric Models

- Integer matrix A of size $(d-1) \times n$
- Map $\psi_A \; : \; (\mathbb{C}^*)^d \longrightarrow (\mathbb{C}^*)^n$ where

$$\psi_A(s, \theta_1, \ldots, \theta_{d-1}) = (s\theta^{a_1},\, s\theta^{a_2},\, \ldots,\, s\theta^{a_n}).$$

- *Toric Variety* V_A defined by image of ψ_A.
- Now, scaling vector $c \in (\mathbb{C}^*)^n$:

$$\psi_A^c(s, \theta_1, \theta_2, \ldots, \theta_{d-1}) \; = \; (c_1 s\theta^{a_1}, c_2 s\theta^{a_2}, \ldots, c_n s\theta^{a_n})$$

- $V_A^c := \overline{\psi_A^c((\mathbb{C}^*)^d)}^Z$ is the scaled toric variety.

Main Question

How does the ML degree of a scaled toric model depend on the scaling?

MRC Likelihood Geometry Group

Carlos Amendola	Nathan Bliss	Isaac Burke
Courtney Gibbons	Martin Helmer	Serkan Hoşten
Evan Nash	Jose Rodriguez	Daniel Smolkin

The Maximum Likelihood Degree of Toric Varieties

`arXiv:1703.02251`

Scaled Twisted Cubic

Example

Consider the scaling vector $c = (1, 3, 3, 1)$. Then for the parametrized scaled twisted cubic:

$$\phi^c(s,t) = (1s, 3st, 3st^2, 1st^3) \subset \Delta_4 \subset \mathbb{R}^4$$

we have that $\mathrm{mldeg}(M_c) =$

Scaled Twisted Cubic

Example

Consider the scaling vector $c = (1, 3, 3, 1)$. Then for the parametrized scaled twisted cubic:

$$\phi^c(s,t) = (1s, 3st, 3st^2, 1st^3) \subset \Delta_4 \subset \mathbb{R}^4$$

we have that $\mathrm{mldeg}(M_c) = 1 < 3 = \deg(M)$.

Birch's Theorem

Theorem (Birch)

Given A for a toric model and a vector of positive counts u with total sum N, the MLE is the unique non-negative solution to the system

$$A\,\hat{p} = \frac{1}{N}\,A\,u$$

with $\hat{p} \in V_A$ (that is, $\hat{p} = \psi_A(\hat{\theta})$).

Birch's Theorem

Theorem (Birch)

Given A for a toric model and a vector of positive counts u with total sum N, the MLE is the unique non-negative solution to the system

$$A\,\hat{p} = \frac{1}{N}\,A\,u$$

with $\hat{p} \in V_A$ (that is, $\hat{p} = \psi_A(\hat{\theta})$).

Remark: It still holds for *scaled* toric models with positive scalings.

Example

Example (Veronese)

$$A = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$$

$$\psi(s, \theta_1, \theta_2) = (s, s\theta_1, s\theta_1^2, s\theta_2, s\theta_1\theta_2, s\theta_2^2)$$

and data vector $u = (1, 3, 5, 7, 9, 2)$.

Example

Example (Veronese)

$$A = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$$

$$\psi(s, \theta_1, \theta_2) = (s, s\theta_1, s\theta_1^2, s\theta_2, s\theta_1\theta_2, s\theta_2^2)$$

and data vector $u = (1, 3, 5, 7, 9, 2)$. Solving the critical equations we obtain the four points

$$(.28887, 1.43166, -1.8931), (.303937, -1.88472, 1.34701)$$

$$(.857893, -.762951, -.718984), (.0863377, 1.63267, 1.51507)$$

Example

Example (Veronese)

$$A = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$$

$$\psi(s, \theta_1, \theta_2) = (s, s\theta_1, s\theta_1^2, s\theta_2, s\theta_1\theta_2, s\theta_2^2)$$

and data vector $u = (1, 3, 5, 7, 9, 2)$. Solving the critical equations we obtain the four points

$$(.28887, 1.43166, -1.8931), (.303937, -1.88472, 1.34701)$$

$$(.857893, -.762951, -.718984), (.0863377, 1.63267, 1.51507)$$

Thus the ML degree is 4 and the MLE is $\hat{\theta} = (.0863377, 1.63267, 1.51507)$.

Scaled Toric Varieties

Example

Let V be the Veronese surface and let $c = (1, 2, 1, 1, 2, 1)$.

$$\psi^{c}(s, \theta_1, \theta_2) = (1s, 2s\theta_1, 1s\theta_1^2, 1s\theta_2, 2s\theta_1\theta_2, 1s\theta_2^2)$$

Scaled Toric Varieties

Example

Let V be the Veronese surface and let $c = (1, 2, 1, 1, 2, 1)$.

$$\psi^c(s, \theta_1, \theta_2) = (1s, 2s\theta_1, 1s\theta_1^2, 1s\theta_2, 2s\theta_1\theta_2, 1s\theta_2^2)$$

$\mathrm{mldeg}(V_c) =$

Scaled Toric Varieties

Example

Let V be the Veronese surface and let $c = (1, 2, 1, 1, 2, 1)$.

$$\psi^c(s, \theta_1, \theta_2) = (1s, 2s\theta_1, 1s\theta_1^2, 1s\theta_2, 2s\theta_1\theta_2, 1s\theta_2^2)$$

$\mathrm{mldeg}(V_c) = 2 < \deg(V_c) = 4$.

Main Result

Theorem (Likelihood Geometry Group)

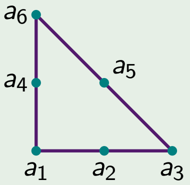
Let $c \in (\mathbb{C}^*)^n$ and let $V \subset \mathbb{P}^{n-1}$ be the toric variety defined by $A \in \mathbb{Z}^{(d-1)\times n}$. Then

- $\mathrm{mldeg}(V_c) \leq \deg(V)$ *and*
- $\mathrm{mldeg}(V_c) < \deg(V)$ *if and only if c is in the hypersurface defined by E_A, the principal A-determinant [GKZ].*

Corollary: For *generic* scalings c, it happens that $\mathrm{mldeg}(V_c) = \deg(V)$

Example

$$A = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$$



- $\Delta_A = \det(C) = \det \begin{pmatrix} c_{00} & c_{10}/2 & c_{01}/2 \\ c_{10}/2 & c_{20} & c_{11}/2 \\ c_{01}/2 & c_{11}/2 & c_{02} \end{pmatrix}.$
- $\Delta_{00,10,20} = c_{10}^2 - 4c_{00}c_{20} \quad \Delta_{00,01,02} = c_{01}^2 - 4c_{00}c_{02}$
  $\Delta_{20,11,02} = c_{11}^2 - 4c_{20}c_{02}$

$$E_A = \det(C)(c_{10}^2 - 4c_{00}c_{20})(c_{01}^2 - 4c_{00}c_{02})(c_{11}^2 - 4c_{20}c_{02})c_{00}c_{20}c_{02}.$$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *)$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *)$ $\qquad$ $\mathrm{mldeg}(X_c) = 3$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *)$ $\qquad \mathrm{mldeg}(X_c) = 3$

$c = (1, 2, 1, 2, *, 1)$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *) \qquad \mathrm{mldeg}(X_c) = 3$

$c = (1, 2, 1, 2, *, 1) \qquad \mathrm{mldeg}(X_c) = 2$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *) \qquad \mathrm{mldeg}(X_c) = 3$

$c = (1, 2, 1, 2, *, 1) \qquad \mathrm{mldeg}(X_c) = 2$

$c = (1, 2, 1, 2, 2, 1)$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *)$ $\quad$ $\mathrm{mldeg}(X_c) = 3$

$c = (1, 2, 1, 2, *, 1)$ $\quad$ $\mathrm{mldeg}(X_c) = 2$

$c = (1, 2, 1, 2, 2, 1)$ $\quad$ $\mathrm{mldeg}(X_c) = 1$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *)$ $\quad \mathrm{mldeg}(X_c) = 3$

$c = (1, 2, 1, 2, *, 1)$ $\quad \mathrm{mldeg}(X_c) = 2$

$c = (1, 2, 1, 2, 2, 1)$ $\quad \mathrm{mldeg}(X_c) = 1$

$c = (1, 4, 1, 6, 6, 6)$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *)$ $\quad$ $\mathrm{mldeg}(X_c) = 3$

$c = (1, 2, 1, 2, *, 1)$ $\quad$ $\mathrm{mldeg}(X_c) = 2$

$c = (1, 2, 1, 2, 2, 1)$ $\quad$ $\mathrm{mldeg}(X_c) = 1$

$c = (1, 4, 1, 6, 6, 6)$ $\quad$ $\mathrm{mldeg}(X_c) = 3$

ML Degree Stratification

Example

$c = (1, 2, 1, *, *, *) \qquad \mathrm{mldeg}(X_c) = 3$

$c = (1, 2, 1, 2, *, 1) \qquad \mathrm{mldeg}(X_c) = 2$

$c = (1, 2, 1, 2, 2, 1) \qquad \mathrm{mldeg}(X_c) = 1$

$c = (1, 4, 1, 6, 6, 6) \qquad \mathrm{mldeg}(X_c) = 3$

Theorem (Likelihood Geometry Group)

Consider the Veronese variety $\mathrm{Ver}(d-1, k)$ *for* $k \leq d-1$ *with scaling given by* $c = (1, 1, \dots, 1, 1)$*. Then* $\mathrm{mldeg}(\mathrm{Ver}(d-1, k)) = k^{d-1}$.

Recall: Homotopy Continuation

- Given F, a polynomial system of equations

$$f_1(x_1, \ldots, x_n) = 0$$

$$f_2(x_1, \ldots, x_n) = 0$$

$$\vdots$$

$$f_n(x_1, \ldots, x_n) = 0.$$

Recall: Homotopy Continuation

- Given F, a polynomial system of equations

$$f_1(x_1, \ldots, x_n) = 0$$
$$f_2(x_1, \ldots, x_n) = 0$$
$$\vdots$$
$$f_n(x_1, \ldots, x_n) = 0.$$

- Choose and solve instead an (easier) polynomial system G based on characteristics of F.

Recall: Homotopy Continuation

- Given F, a polynomial system of equations

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ f_2(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n) &= 0. \end{aligned}$$

- Choose and solve instead an (easier) polynomial system G based on characteristics of F.
- Form the homotopy system $H(x,t) = (1-t) \cdot F(x) + t \cdot G(x)$

Recall: Homotopy Continuation

- Given F, a polynomial system of equations

$$
\begin{aligned}
f_1(x_1, \ldots, x_n) &= 0 \\
f_2(x_1, \ldots, x_n) &= 0 \\
&\vdots \\
f_n(x_1, \ldots, x_n) &= 0.
\end{aligned}
$$

- Choose and solve instead an (easier) polynomial system G based on characteristics of F.
- Form the homotopy system $H(x, t) = (1 - t) \cdot F(x) + t \cdot G(x)$
- Use predictor-corrector methods to track the numerical solutions as t moves from $t = 1$ to $t = 0$.

Homotopy Tracking

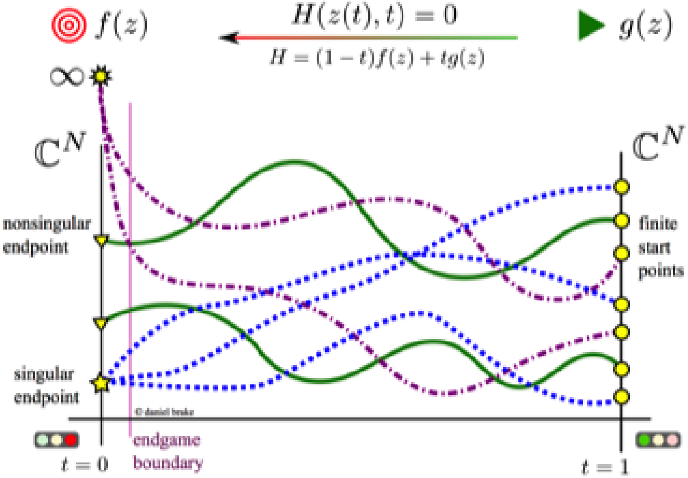


Figure: Homotopy Continuation Illustration (Dani Brake)

Homotopy Tracking

Theorem (Likelihood Geometry Group)

Fix a generic data vector u with positive entries. Let c_{win} and c_{stat} be scalings with positive entries.

Homotopy Tracking

Theorem (Likelihood Geometry Group)

Fix a generic data vector u with positive entries. Let c_{win} and c_{stat} be scalings with positive entries.Consider the homotopy with target system the critical likelihood equations for the model $V_A^{c_{stat}}$ and start system the ones for $V_A^{c_{win}}$, with data vector u.

Homotopy Tracking

Theorem (Likelihood Geometry Group)

Fix a generic data vector u with positive entries. Let c_{win} and c_{stat} be scalings with positive entries.Consider the homotopy with target system the critical likelihood equations for the model $V_A^{c_{stat}}$ and start system the ones for $V_A^{c_{win}}$, with data vector u.
Let $\hat{\theta}_{win}$ and $\hat{\theta}_{stat}$ be the respective MLEs and let γ denote the path of the homotopy whose start point (at $t = 1$) corresponds to $\hat{\theta}_{win}$.

Homotopy Tracking

Theorem (Likelihood Geometry Group)

Fix a generic data vector u with positive entries. Let c_{win} and c_{stat} be scalings with positive entries.Consider the homotopy with target system the critical likelihood equations for the model $V_A^{c_{stat}}$ and start system the ones for $V_A^{c_{win}}$, with data vector u.
Let $\hat{\theta}_{win}$ and $\hat{\theta}_{stat}$ be the respective MLEs and let γ denote the path of the homotopy whose start point (at $t = 1$) corresponds to $\hat{\theta}_{win}$. Then, the endpoint of γ (at $t = 0$) is $\hat{\theta}_{stat}$.

Proof Sketch

- By *Birch's Theorem*, a homotopy between the two systems is given by

$$H(\theta, t) := t\left(A\hat{p}_{stat} - \frac{1}{N}Au\right) + (1-t)\left(A\hat{p}_{win} - \frac{1}{N}Au\right)$$

Proof Sketch

- By *Birch's Theorem*, a homotopy between the two systems is given by

$$H(\theta, t) := t\left(A\hat{p}_{stat} - \frac{1}{N}Au\right) + (1-t)\left(A\hat{p}_{win} - \frac{1}{N}Au\right)$$

- This simplifies to $A \cdot (\hat{p}_{c(t)} - \frac{1}{N}u)$ where $c(t) = tc_{stat} + (1-t)c_{win}$

Proof Sketch

- By *Birch's Theorem*, a homotopy between the two systems is given by

$$H(\theta, t) := t\left(A\hat{p}_{stat} - \frac{1}{N}Au\right) + (1-t)\left(A\hat{p}_{win} - \frac{1}{N}Au\right)$$

- This simplifies to $A \cdot (\hat{p}_{c(t)} - \frac{1}{N}u)$ where $c(t) = tc_{stat} + (1-t)c_{win}$
- For positive real c_{win}, c_{stat}, we have $c(t) > 0$ for any $t \in [0,1]$. Thus by *Birch's Theorem* there is exactly one positive real solution to the system at every point along the homotopy path.

Proof Sketch

- By *Birch's Theorem*, a homotopy between the two systems is given by

$$H(\theta, t) := t\left(A\hat{p}_{stat} - \frac{1}{N}Au\right) + (1-t)\left(A\hat{p}_{win} - \frac{1}{N}Au\right)$$

- This simplifies to $A \cdot (\hat{p}_{c(t)} - \frac{1}{N}u)$ where $c(t) = tc_{stat} + (1-t)c_{win}$
- For positive real c_{win}, c_{stat}, we have $c(t) > 0$ for any $t \in [0,1]$. Thus by *Birch's Theorem* there is exactly one positive real solution to the system at every point along the homotopy path.
- Left to show tracking paths do not intersect (we show the Jacobian matrix of the system has always full rank)

□

Possible Applications

- In practice, a statistical toric model will come with a specified scaling c_{stat}.

Possible Applications

- In practice, a statistical toric model will come with a specified scaling c_{stat}.
- Knowing how scaling vectors c affect the ML degree of a particular toric model V_A allows us to find a convenient c_{win} (e.g. such that the model has *low* ML degree).

Possible Applications

- In practice, a statistical toric model will come with a specified scaling c_{stat}.
- Knowing how scaling vectors c affect the ML degree of a particular toric model V_A allows us to find a convenient c_{win} (e.g. such that the model has *low* ML degree).
- By the Theorem, we can now find the MLE $\hat{\theta}_{win}$ and track its unique homotopy path to find the original MLE of interest $\hat{\theta}_{stat}$.

Application Example

Example (Veronese revisited)

Recall

$$A = \begin{bmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{bmatrix},$$

with $u = (1, 3, 5, 7, 9, 2)$. Here $c_{stat} = (1, 1, 1, 1, 1, 1)$.

Application Example

Example (Veronese revisited)

Recall

$$A = \begin{bmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{bmatrix},$$

with $u = (1, 3, 5, 7, 9, 2)$. Here $c_{stat} = (1, 1, 1, 1, 1, 1)$. By choosing $c_{win} = (1, 2, 1, 2, 2, 1)$, the ML degree drops to $\mathbf{1}$.

Application Example

Example (Veronese revisited)

Recall

$$A = \begin{bmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{bmatrix},$$

with $u = (1, 3, 5, 7, 9, 2)$. Here $c_{stat} = (1, 1, 1, 1, 1, 1)$. By choosing $c_{win} = (1, 2, 1, 2, 2, 1)$, the ML degree drops to $\mathbf{1}$. Computing the unique critical point we obtain the MLE $\hat{\theta}_{win} = (.0493827, 1.83333, 1.66667)$. Tracking this point in the homotopy we arrive at the point $\hat{\theta}_{track} = (.0863377, 1.63267, 1.51507)$.

Application Example

Example (Veronese revisited)

Recall

$$A = \begin{bmatrix} 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{bmatrix},$$

with $u = (1, 3, 5, 7, 9, 2)$. Here $c_{stat} = (1, 1, 1, 1, 1, 1)$. By choosing $c_{win} = (1, 2, 1, 2, 2, 1)$, the ML degree drops to $\mathbf{1}$. Computing the unique critical point we obtain the MLE $\hat{\theta}_{win} = (.0493827, 1.83333, 1.66667)$. Tracking this point in the homotopy we arrive at the point $\hat{\theta}_{track} = (.0863377, 1.63267, 1.51507)$. This coincides with the MLE $\hat{\theta}_{stat}$ computed before.

Success story

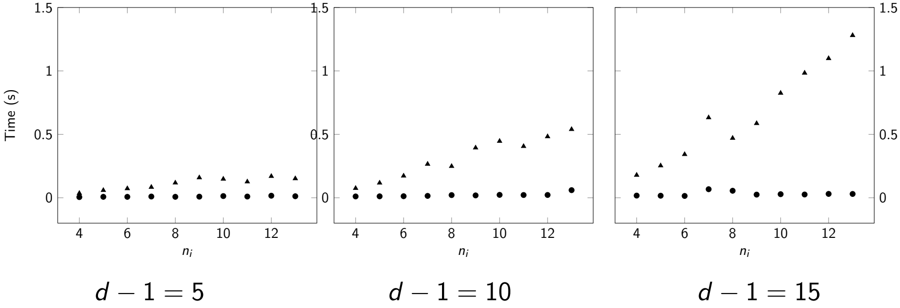


Figure: Running times of iterative proportional scaling (triangles) versus path tracking (circles) on rational normal scrolls. Average of 7 trials.

Advertisement: Check out poster at SIAM Applied Algebraic Geometry (Monday PP1 Welcome Reception and Poster Session) presented by Evan Nash:

Maximum Likelihood Estimate Homotopy Tracking for Toric Models

Advertisement: Check out poster at SIAM Applied Algebraic Geometry (Monday PP1 Welcome Reception and Poster Session) presented by Evan Nash:

Maximum Likelihood Estimate Homotopy Tracking for Toric Models

Advertisement: *Algebraic Statistics Day* on August 11 at MPI Leipzig!

Advertisement: Check out poster at SIAM Applied Algebraic Geometry (Monday PP1 Welcome Reception and Poster Session) presented by Evan Nash:

Maximum Likelihood Estimate Homotopy Tracking for Toric Models

Advertisement: *Algebraic Statistics Day* on August 11 at MPI Leipzig!

THANK YOU!