Chapter 7: Likelihood Inference

Carlos Enrique Améndola Cerón

Introduction to Algebraic Statistics Course

May 25, 2018
Berlin

Maximum Likelihood Estimation

Recall from Chapter 5:

- (Def 5.3.5) Likelihood function for a model M_Θ with data D: $L(\theta|D)$ $(= \ p_\theta(D)$ or $f_\theta(D))$.
- MLE $\hat{\theta}$ maximizes the (log-)likelihood function:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta|D)$$

- (Def 7.1.1) The score equations are obtained by setting the gradient of the log-likelihood to zero: $\frac{\partial}{\partial \theta_i}\ell(\theta|D) = 0$ for $i = 1, \ldots, d$.
- In the discrete case $p : \Theta \to \Delta_{r-1}$: for i.i.d. data $X^{(1)}, \ldots, X^{(n)}$ summarized by the vector of counts $u \in \mathbb{N}^r$, we have

$$\ell(\theta|u) = \sum_{j=1}^{r} u_j \log p_j(\theta).$$

The ML degree

- $\ell(\theta|u) = \sum_{j=1}^{r} u_j \log p_j(\theta)$, hence score equations are *rational*:

$$\sum_{j=1}^{r} \frac{u_j}{p_j} \frac{\partial p_j}{\partial \theta_i}(\theta) = 0 \quad i = 1 \ldots, d.$$

Theorem (Thm 7.1.2, Def 7.1.4)

Let $p : \Theta \to \Delta_{r-1}$. For generic data, the number of (complex) solutions to the score equations is independent of u. We call this the ML degree of the parametric discrete statistical model $M_\Theta \subset \Delta_{r-1}$.

- ML degree measures the complexity of the ML estimation problem.
- ML degree is 1 $\iff$ the MLE is a rational function of the data.

Example (Twisted Cubic Model)

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

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

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

The ML degree is 3.

- Recall (Prop 5.3.7) the Gaussian model log-likelihood $\ell(\mu, \Sigma | \bar{X}, S)$:

$$-\frac{n}{2}(\log\det\Sigma + m\log 2\pi) - \frac{n}{2}\mathrm{tr}(S\Sigma^{-1}) - \frac{n}{2}(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu).$$

Example (Prop 7.1.6)

Let $\Theta = \Theta_1 \times Id_m \subset \mathbb{R}^m \times PD_m$ for a Gaussian statistical model. Then the maximum likelihood estimation for Θ is equivalent to the least-squares point on Θ_1. In this case, $\mathrm{ML\ degree} = \#$ critical points of $||\bar{X}-\mu||_2^2$, known as the ED degree of Θ_1.

- (Prop 7.1.9) Let $\Theta = \mathbb{R}^m \times \Theta_2 \subset \mathbb{R}^m \times PD_m$ for a Gaussian statistical model. Then ML estimation gives $\hat{\mu} = \bar{X}$ and reduces to maximizing $-\frac{n}{2}\log\det\Sigma - \frac{n}{2}\mathrm{tr}(S\Sigma^{-1})$.

Example (Ex 7.1.11 Gaussian Marginal Independence)

Let $\Theta = \mathbb{R}^m \times \Theta_2$ where $\Theta_2 = \{\Sigma \in PD_4 | \sigma_{12} = \sigma_{21} = 0, \sigma_{34} = \sigma_{43} = 0\}$. The marginal independence constraints are $X_1 \perp\!\!\!\perp X_2$ and $X_3 \perp\!\!\!\perp X_4$. The ML degree is found to be 17.

Likelihood Geometry

Definition (ML degree of a variety)

Let $V \subset \mathbb{P}^{r-1}$ be an irreducible projective variety over $\mathbb{C}$, $u \in \mathbb{N}^r$ and

$$L_u(p) = \frac{p_1^{u_1} p_2^{u_2} \cdots p_r^{u_r}}{(p_1 + \cdots + p_r)^{u_1 + \cdots + u_r}}.$$

ML degree of V is the number of (complex) critical points for generic u of $L_u(p)$ on $V_{reg} \setminus \mathcal{H}$, where $\mathcal{H} = \{p \in \mathbb{P}^{r-1} : p_1 \cdots p_r(p_1 + \cdots + p_r) = 0\}$.

- If $I(V) = \langle f_1, f_2, \ldots, f_k \rangle$, use *Lagrange multipliers* to optimize L.
- (Thm 7.2.9) Huh (2013): the ML degree of a smooth very affine variety (of the form $V \cap (\mathbb{C}^*)^r$ where $V \subset \mathbb{C}^r$ variety) is $\pm\chi_{\mathrm{top}}(\cdot)$.
- (Theorem 7.2.13) Huh (2014): Characterization of ML degree 1 varieties as *A*-discriminants [GKZ] (via *Horn uniformization*).

ML in Exponential Families

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.

Discrete and Gaussian exponential families revisited

Corollary (Birch's Theorem, Cor 7.3.9)

The MLE in the log-linear model $\mathcal{M}_{A,h}$ given the data u is the unique solution, if it exists, to the equations

$$Au = n\,A\hat{p} \qquad \text{and} \qquad \hat{p} \in \mathcal{M}_{A,h}$$

Inspires algorithms for computing MLE: Iterative Proportional Scaling (IPS)

Corollary (Cor 7.3.10)

Let $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^m$ i.i.d. samples from the Gaussian exponential family parametrized by $(\mu, \Sigma) \in \mathbb{R}^m \times \mathcal{M}_{L^{-1}}$ (L linear space such that $L \cap PD_m \neq \emptyset$). The MLE is $(\bar{X}, \hat{S})$ where $\hat{S}$ is the unique solution (if it exists) to the equations

$$\pi(S) = \pi(\hat{S}) \qquad \text{and} \qquad \hat{S} \in \mathcal{M}_{L^{-1}}$$

where π denotes the orthogonal projection onto L.

Exercise (cf. Ex. 7.2)

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

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

- 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.