More on Maximum Likelihood Estimation

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

1. Consistent Roots of the Likelihood Equations
2. Approximate Solution of the Likelihood Equations
3. The Multiparameter Case
4. Misspecified Models and Likelihood

Maximum Likelihood Estimators

Recall our definition of a maximum likelihood estimator:

Definition (Maximum Likelihood Estimators)

Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a random sample from F_θ, and suppose that $\hat{\theta}$ is such that

$$L(\hat{\theta}) \geq L(\theta), \quad \forall\ \theta \in \Theta.$$

Then $\hat{\theta}$ is called *a maximum likelihood estimator of* θ.

We saw that, under regularity conditions, the distribution of a consistent sequence of MLEs converges weakly to the normal distribution centred around the true parameter value.

- Consistent likelihood equation roots
- Newton-Raphson and "one-step" estimators
- The multivariate parameter case
- What happens if the model has been mis-specified?

Consistent Roots of the Likelihood Equations

Consistent Likelihood Roots

Theorem

Let $\{f(\cdot;\theta)\}_{\theta\in\mathbb{R}}$ be an identifiable parametric class of densities (frequencies) and let $X_1, ..., X_n$ be iid random variables each having density $f(x;\theta_0)$. If the support of $f(\cdot;\theta)$ is independent of θ,

$$\mathbb{P}[L(\theta_0|X_1, ..., X_n) > L(\theta|X_1, ..., X_n)] \overset{n\to\infty}{\longrightarrow} 1$$

for any fixed $\theta \neq \theta_0$.

- Therefore, with high probability, the likelihood of the true parameter exceeds the likelihood of any other choice of parameter, provided that the sample size is large.
- Hints that extrema of $L(\theta; \boldsymbol{X})$ should have something to do with θ_0 (even though we saw that without further assumptions a maximizer of L is not necessarily consistent).

Proof.

Notice that

$$L(\theta_0|\boldsymbol{X}_n) > L(\theta|\boldsymbol{X}_n) \iff \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right] < 0$$

By the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right] \xrightarrow{p} \mathbb{E}\log\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right] = -KL(f_\theta \| f_{\theta_0})$$

But we have seen that the KL-divergence is zero only at θ_0 and positive everywhere else.

□

Consistent Sequences of Likelihood Roots

Theorem (Cramér)

Let $\{f(\cdot;\theta)\}_{\theta\in\mathbb{R}}$ be an identifiable parametric class of densities/frequencies, and Θ open. Let $X_1, ..., X_n$ be iid random variables each having density $f(x;\theta_0)$. Assume that the the support of $f(\cdot;\theta)$ is independent of θ and that $f(x;\theta)$ is differentiable with respect to θ for (almost) all x. Then, there exists a sequence of random variables ξ_n such that

$$\ell'(X_1, ..., X_n; \xi_n) = 0, \quad \forall\, n \geq 1$$

and

$$\xi_n \xrightarrow{p} \theta_0.$$

- In other words, there exists a sequence of roots of the likelihood equations that is consistent for θ_0.
- In general ξ_n is not a statistic (and so not an estimator), since $\xi_n = g(X_1, ..., X_n; \theta_0)$ – we need to know the true θ_0 in order to choose which of the likelihood roots to select as our ξ_n for a given sample $(X_1, ..., X_n)$.

Proof.

Let $\alpha > 0$ be sufficiently small so that $(\theta_0 - \alpha, \theta_0 + \alpha) \subset \Theta$. Define the set

$$S_n(\alpha, \theta_0) := \left\{ \mathbf{x} \in \mathbb{R}^n : \ell(\mathbf{x}; \theta_0) > \ell(\mathbf{x}; \theta_0 - \alpha) \quad \& \quad \ell(\mathbf{x}; \theta_0) > \ell(\mathbf{x}; \theta_0 + \alpha) \right\}$$

If $\mathbf{x} \in S_n(\alpha, \theta_0)$, there exists at least one local maximum of $\ell(\boldsymbol{x}; \theta)$ in $(\theta_0 - \alpha, \theta_0 + \alpha)$, and hence at least one point $t \in (\theta_0 - \alpha, \theta_0 + \alpha)$ in that interval such that $\ell'(\mathbf{x}; t) = 0$. Choose $\tilde{\xi}(\mathbf{x}, \alpha, \theta_0)$ to be the local minimum closest to θ_0 when $\mathbf{x} \in S_n(\alpha, \theta_0)$ or zero if $\mathbf{x} \notin S_n(\alpha, \theta_0)$.

By our previous theorem we know that there exists[a] $\alpha_n \downarrow 0$ such that $\mathbb{P}_{\theta_0}[S_n(\alpha_n, \theta_0)] \overset{n \to \infty}{\longrightarrow} 1$. Define $\xi_n = \tilde{\xi}(\mathbf{x}, \alpha_n, \theta_0)$.

Now pick $\delta > 0$. Then, for n sufficiently large (so that $\alpha_n < \delta$) we have

$$\mathbb{P}_{\theta_0}[|\xi_n - \theta_0| < \delta] \geq \mathbb{P}_{\theta_0}[|\xi_n - \theta_0| < \alpha_n] \geq \mathbb{P}_{\theta_0}[S_n(\alpha_n)] \to 1$$

because $\mathbf{x} \in S_n(\alpha_n) \implies |\xi_n - \theta_0| < \alpha_n$, and the proof is complete. □

[a] Exercise: show this using the same trick as with the Ky-Fan definition of $\overset{p}{\to}$

Corollary (Consistency of Unique Solutions)

Under the assumptions of the previous theorem, if the likelihood equation has a unique root ξ_n for each n and all $\mathbf{x}$, then ξ_n is a valid estimator and is consistent for θ_0.

- The statement remains true if the uniqueness requirement is substituted with the requirement that the probability of multiple roots tends to zero as $n \to \infty$.
- Notice that the statement does not claim that the root corresponds to a maximum: it merely requires that we have a root.
- On the other hand, even when the root is unique, the corollary says nothing about its properties for finite n.

Example (Minimum Likelihood Estimation)

Let X take the values $0, 1, 2$ with probabilities $6\theta^2 - 4\theta + 1$, $\theta - 2\theta^2$ and $3\theta - 4\theta^2$ ($\theta \in (0, 1/2)$). Then, the likelihood equation has a unique root for all x, which is a minimum for $x = 0$ and a maximum for $x = 1, 2$.

Consistent Sequences of Likelihood Roots

- Cramér does not tell us *which* root to choose, so not useful in practice
- The easiest case is when the root is unique!
- Otherwise, we need some "external help" (non-MLE help)...

Fortunately, some "good" estimator is already available, then...

Lemma

Let α_n be any consistent sequence of estimators for the parameter θ. For each n, let θ_n^* denote the root of the likelihood equations that is closest to α_n. Then, under the assumptions of Cramér's theorem, $\theta_n^* \to \theta$.

Exercise: prove the lemma.

- Therefore, when the likelihood equations do not have a single root, we may still choose a root based on some estimator that is readily available
  - ↪ Only require that the estimator used is consistent
  - ↪ Often the case with Plug-In or MoM estimators

Very often, the roots will not be available in closed form. In these cases, an iterative approach will be required to approximate the roots

Approximate Solution of the Likelihood Equations

The Newton-Raphson Algorithm

We wish to solve the equation

$$\ell'(\theta) = 0$$

Supposing that $\tilde{\theta}$ is close to a root (perhaps is a consistent estimator),

$$0 = \ell'(\hat{\theta}) \simeq \ell'(\tilde{\theta}) + (\hat{\theta} - \tilde{\theta})\ell''(\tilde{\theta})$$

By using a second-order Taylor expansion. This suggests

$$\hat{\theta} \simeq \tilde{\theta} - \frac{\ell'(\tilde{\theta})}{\ell''(\tilde{\theta})}$$

The procedure can then be iterated by replacing $\tilde{\theta}$ by the right hand side of the above relation. In principle, each iteration improves the finite sample accuracy of our estimator – but in terms of asymptotic behaviour, a single iteration suffices!

Construction of Asymptotically MLE-like Estimators

Theorem

Suppose that assumptions (A1)-(A6) hold and let $\tilde{\theta}_n$ be a consistent estimator of θ_0 such that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability. Then, the sequence of estimators

$$\delta_n = \tilde{\theta}_n - \frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}$$

satisfies

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta)/J(\theta)^2).$$

- Therefore, with a single Newton-Raphson step, we may obtain an estimator that, asymptotically, behaves like a consistent MLE.
  - ↪ Provided that we have a $\sqrt{n}-$consistent estimator!
- "One-step" not necessarily behaving like an MLE for finite n!
- Note that the one step δ_n satisfies the conditions of theorem (is consistent and bounded in probability). Hence iterating to get $\zeta_n = \delta_n - \ell'(\delta_n)/\ell''(\delta_n)$ also gives us the same conclusion.

Proof.

We Taylor expand around the true value, θ_0,

$$\ell'(\tilde{\theta}_n) = \ell'(\theta_0) + (\tilde{\theta}_n - \theta_0)\ell''(\theta_0) + \tfrac{1}{2}(\tilde{\theta}_n - \theta_0)^2\ell'''(\theta_n^*)$$

with θ_n^* between θ_0 and $\tilde{\theta}_n$. Substituting this expression into the definition of δ_n yields

$$\begin{aligned}\sqrt{n}(\delta_n - \theta_0) &= \frac{(1/\sqrt{n})\ell'(\theta_0)}{-(1/n)\ell''(\tilde{\theta}_n)} + \sqrt{n}(\tilde{\theta}_n - \theta_0) \times \\ &\quad \times \left[1 - \frac{\ell''(\theta_0)}{\ell''(\tilde{\theta}_n)} - \frac{1}{2}(\tilde{\theta}_n - \theta_0)\frac{\ell'''(\theta_n^*)}{\ell''(\tilde{\theta}_n)}\right]\end{aligned}$$

Exercise

Use CLT/LLN/Slutsky to complete the proof. Hint: by Taylor expansion,

$$\tfrac{1}{n}\ell''(\tilde{\theta}_n) = \tfrac{1}{n}\textstyle\sum_i \ell''(X_i;\tilde{\theta}_n) = \tfrac{1}{n}\sum_i \ell''(X_i;\theta_0) + (\tilde{\theta}_n - \theta_0)\tfrac{1}{n}\sum_i \ell'''(X_i;\theta_0)$$

The Multiparameter Case

The Multiparameter Case

$\rightarrow$ Extension of asymptotic results to multiparameter models easy under similar assumptions, but notationally cumbersome.
$\rightarrow$ Same ideas: the MLE will be a zero of the likelihood equations

$$\sum_{i=1}^{n} \nabla \ell(X_i; \boldsymbol{\theta}) = 0$$

A Taylor expansion can be formed

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ell(X_i; \boldsymbol{\theta}) + \left(\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(X_i; \boldsymbol{\theta}_n^*) \right) \sqrt{n}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$$

Under regularity conditions we should have:

- $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ell(X_i; \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(0, \text{Cov}[\nabla \ell(X_i; \boldsymbol{\theta})])$
- $\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(X_i; \boldsymbol{\theta}_n^*) \xrightarrow{p} \mathbb{E}[\nabla^2 \ell(X_i; \boldsymbol{\theta})]$

The Multiparameter Case

Regularity Conditions

(B1) The parameter space $\Theta \in \mathbb{R}^p$ is open.

(B2) The support of $f(\cdot|\boldsymbol{\theta})$, $\text{supp} f(\cdot|\boldsymbol{\theta})$, is independent of $\boldsymbol{\theta}$

(B3) All mixed partial derivatives of ℓ w.r.t. θ up to degree 3 exist and are continuous.

(B4) $\mathbb{E}[\nabla \ell(X_i; \boldsymbol{\theta})] = 0 \ \forall \boldsymbol{\theta}$ and $\text{Cov}[\nabla \ell(X_i; \boldsymbol{\theta})] =: I(\boldsymbol{\theta}) \succ 0 \ \forall \boldsymbol{\theta}$.

(B5) $-\mathbb{E}[\nabla^2 \ell(X_i; \boldsymbol{\theta})] =: J(\boldsymbol{\theta}) \succ 0 \ \forall \boldsymbol{\theta}$.

(B6) $\exists \delta > 0$ s.t. $\forall \boldsymbol{\theta} \in \Theta$ and for all $1 \leq j, k, l \leq p$,

$$\left| \frac{\partial}{\partial \theta_j \partial \theta_k \partial \theta_l} \ell(x; \boldsymbol{u}) \right| \leq M_{jkl}(x)$$

for $\|\boldsymbol{\theta} - \boldsymbol{u}\| \leq \delta$ with M_{jkl} such that $\mathbb{E}[M_{jkl}(X_i)] < \infty$.

- The interpretation of the conditions is the same as with the one-dimensional case

The Multiparameter Case

Theorem (Asymptotic Normality of the MLE)

Let $X_1, ..., X_n$ be iid random variables with density (frequency) $f(x; \boldsymbol{\theta})$, satisfying conditions (B1)-(B6). If $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(X_1, ..., X_n)$ is a consistent sequence of MLE estimators, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(0, J^{-1}(\boldsymbol{\theta}) I(\boldsymbol{\theta}) J^{-1}(\boldsymbol{\theta}))$$

- The theorem remains true if each X_i is a random vector
- The proof mimics that of the one-dimensional case

Misspecified Models and Likelihood

Misspecification of Models

- Statistical models are typically mere approximations to reality
- George P. Box: "*all models are wrong, but some are useful*"

As worrying as this may seem, it may not be a problem in practice.

- Often the model is wrong, but is "close enough" to the true situation
- Even if the model is wrong, the parameters often admit a fruitful interpretation in the context of the problem.

Example

Let $X_1, ..., X_n$ be iid Exponential(λ) r.v.'s but we have modelled them as having the following two parameter density

$$f(x|\alpha, \theta) = \frac{\alpha}{\theta}\left(1 + \frac{x}{\theta}\right)^{-(\alpha+1)}, \quad x > 0$$

with α and θ positive unknown parameters to be estimated.

Example (cont'd)

- Notice that the exponential distribution is not a member of this parametric family.
- However, letting $\alpha, \theta \to \infty$ such that $\alpha/\theta \to \lambda$, we have

$$f(x|\alpha, \theta) \to \lambda \exp(-\lambda x)$$

Thus, we may *approximate* the true model from within this class. Reasonable $\hat{\alpha}$ and $\hat{\lambda}$ will yield a density "close" to the true density.

Example

Let $X_1, ..., X_n$ be independent random variables with variance σ^2 and mean

$$\mathbb{E}[X_i] = \alpha + \beta t_i$$

If we assume that the X_i are normal when they are in fact not, the MLEs of the parameters α, β, σ^2 remain good (in fact optimal in a sense) for the true parameters (Gauss-Markov theorem).

Misspecified Models and Likelihood

The Framework

- $X_1, ..., X_n$ are iid r.v.'s with distribution F
- We build a MLE assuming that the X_i admit a density in $\{f(x;\theta)\}_{\theta\in\Theta}$.
- The true distribution F does not correspond to any of the $\{f_\theta\}$

Let $\hat{\theta}_n$ be a root of the likelihood equation,

$$\sum_{i=1}^{n} \ell'(X_i;\hat{\theta}_n) = 0$$

where the log-likelihood $\ell(\theta)$ is w.r.t. $f(\cdot|\theta)$.

- What exactly is $\hat{\theta}_n$ estimating?
- What is the behaviour of the sequence $\{\hat{\theta}_n\}_{n\geq 1}$ as $n \to \infty$?

Misspecified Models and Likelihood

Consider the functional parameter $\theta(F)$ defined by

$$\int_{-\infty}^{+\infty} \ell'(x; \theta(F)) dF(x) = 0$$

Then, the plug-in estimator of $\theta(F)$ when using the edf $\hat{F}_n$ as an estimator of F is given by solving

$$\int_{-\infty}^{+\infty} \ell'(x; \theta(\hat{F}_n)) d\hat{F}_n(x) = 0 \iff \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0$$

so that the MLE is a plug-in estimator of $\theta(F)$.

Model Misspecification and the Likelihood

Theorem

Let $X_1, ..., X_n \overset{iid}{\sim} F$ *and let* $\hat{\theta}_n$ *be a random variable solving the equations* $\sum_{i=1}^{n} \ell'(X_i; \theta) = 0$ *for* θ *in the open set* Θ. *If*

(a) *For each* x, *the likelihood* $\theta \to \ell(x, \theta)$ *is strictly concave.*

Furthermore, $\theta \to E(\ell(X, \theta))$ *has a maximum:* $\theta(F)$.

(b) $I(F) := \int_{-\infty}^{+\infty} [\ell'(x; \theta(F))]^2 dF(x) < \infty$

(c) $J(F) := -\int_{-\infty}^{+\infty} \ell''(x; \theta(F)) dF(x) < \infty$

(d) $|\ell'''(x; t)| \leq M(x)$ *for* $t \in (\theta(F) - \delta, \theta(F) + \delta)$, *some* $\delta > 0$ *and* $\int_{-\infty}^{+\infty} M(x) dF(x) < \infty$

Then

$$\hat{\theta}_n \xrightarrow{p} \theta(F)$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta(F)) \xrightarrow{d} \mathcal{N}(0, I(F)/J^2(F))$$

Proof

First, let us study $\theta(F)$.

$$\theta \to E(\ell(X, \theta))$$

is strictly concave and doesn't asymptote to a finite value. Thus, $\theta(F)$ exists and is unique.

For any value of the data, $\hat{\theta}_{ML}$ is unique (if it exists).
Our proof proves simultaneously the consistency and the Gaussian limit behavior, by performing a Taylor expansion around $\theta(F)$ of the empirical log-likelihood.

Proof

Let $\ell_n(\theta) = \sum_{i=1}^n \ell(X_i, \theta)$ and $M_n = \sum_{i=1}^n M(X_i)$.
We have the following results:

1. Central limit theorem:

$$\frac{1}{\sqrt{n}} \ell_n'(\theta(F)) \xrightarrow{p} N(0, I(F))$$

2. Law of large numbers:

$$-\frac{1}{n} \ell_n''(\theta(F)) \xrightarrow{p} J(F)$$

3. Law of large numbers:

$$\frac{1}{n} M_n \xrightarrow{p} E(M(X))$$

Proof

Furthermore, for $t \in [\theta(F) - \delta, \theta(F) + \delta]$, $\ell_n'(t)$ is bounded by the following Taylor expansion:

$$|\ell_n'(t) - \ell_n'(\theta(F)) - \ell_n''(\theta(F))(t - \theta(F))| \leq M_n \frac{(t - \theta(F))^2}{2}$$

From the intermediate value theorem (draw it !), $\hat{\theta}_n$ is between the roots of the following two quadratics (if they are in the δ ball around $\theta(F)$):

$$\ell_n'(\theta(F)) + \ell_n''(\theta(F))(t - \theta(F)) + M_n \frac{(t - \theta(F))^2}{2}$$

$$\ell_n'(\theta(F)) + \ell_n''(\theta(F))(t - \theta(F)) - M_n \frac{(t - \theta(F))^2}{2}$$

Critically, these polynomials are not guaranteed to have roots !

Proof

Let's study one:

$$\ell'_n(\theta(F)) + \ell''_n(\theta(F))(t - \theta(F)) + M_n \frac{(t - \theta(F))^2}{2}$$

Its determinant is:

$$\Delta = \left[\ell''_n(\theta(F))\right]^2 - 4\ell'_n(\theta(F))M_n$$

$$\frac{1}{n^2}\Delta = \left[\frac{\ell''_n(\theta(F))}{n}\right]^2 - 4\frac{\ell'_n(\theta(F))}{n}\frac{M_n}{n}$$

$$\frac{1}{n^2}\Delta \xrightarrow{p} J(F)^2 > 0$$

Thus, $\Delta > 0$ occurs with probability tending to 1: this quadratic is guaranteed to have two roots !

Proof

As an exercise, prove that in the limit $n \to \infty$, the roots are approximately:

$$\theta^\star \approx \theta(F) - \frac{l_n'(\theta(F))}{l_n''(\theta(F))}$$

We finally have:

$$\sqrt{n}(\hat{\theta}_{ML} - \theta(F)) \approx -\frac{\sqrt{n} l_n'(\theta(F))}{l_n''(\theta(F))} \xrightarrow{p} N(0, I(F)/J(F)^2)$$

- The result extends immediately to the multivariate parameter case.
- The proof is essentially identical to MLE asymptotics proof.
- Assumption (a) gives us consistency.
- We have assumed our model has log-concave likelihoods ! This strong assumption can be replaced by any set of assumptions yielding consistency.

Model Misspecification and the Likelihood

What is the interpretation of the parameter $\theta(F)$ in the misspecified setup?

Suppose that F has density (frequency) g and assume that integration/differentiation may be interchanged:

$$\int_{-\infty}^{+\infty} \frac{d}{d\theta} \log f(x;\theta) dF(x) = 0 \iff \frac{d}{d\theta} \int_{-\infty}^{+\infty} \log f(x;\theta) dF(x) = 0$$

$$\iff \frac{d}{d\theta} \left[\int_{-\infty}^{+\infty} \log f(x;\theta) dF(x) - \int_{-\infty}^{+\infty} \log g(x) dF(x) \right] = 0$$

$$\iff \frac{d}{d\theta} KL(g(x) \| f(x;\theta)) = 0$$

- We are minimizing the KL-distance between the true model F and our model.
- Hence we may intuitively think of the $\theta(F)$ as the element of Θ for which f_θ is "closest" to F in the KL-sense.

Summary

Last week, we talked about the MLE which is asymptotically Gaussian if it is consistent. Consistency proved slightly hard to study.

This week, we showed that we can bootstrap from any other $\sqrt{n}$ consistent estimator $\hat{\theta}$: by adding a small Newton-Raphson correction to it, we obtain a true estimator that is $\sqrt{n}$ consistent and asymptotically Gaussian.
This is how you use the MLE in complicated cases !

We also considered what happens when the true model isn't inside our parametric family:

- We are trying to infer the best approximation of the truth inside our model class, parameterized by $\theta(F)$.
- Up to possible issues of consistency (which we dealt with by assuming concavity), the MLE correctly recovers $\theta(F)$ and is asymptotically Gaussian.