MSF100
Statistical Inference Principles - Lecture 5

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

January 31, 2012

1 Properties of the MLE

We will continue with the properties of the MLE in this lecture. Let us first review what we already know. Last week we showed that

$$E_\theta[S(\theta)] = 0, \quad Var_\theta[S(\theta)] = -E_\theta[\frac{d^2}{d\theta^2}\log L(\theta)]$$

under regularity conditions (smoothness of pdf f_θ). The last quantity is called the *Fisher Information* or *expected* information, $FI(\theta)$, and is the expected value of the curvature of the loglikelihood function. The *observed* information $I(\theta) = -\frac{d^2}{d\theta^2}l(\theta)$ says something about the uncertainty about θ given *this* sample.

We have already seen that the *exponential family* of distributions are easy to work with and show results for (like UMVUE etc). Here is another fact about exponential families. It turns out that the observed and expected information coincide at $\hat{\theta}$ (the MLE) and in the special case of *canonical parameters* the entire function $I(\theta)$ and $FI(\theta)$ coincide!

Here's the proof.

- The pdf of an exponential family can be written as $p_\theta(x) = exp(w(\theta)T(x) - A(\theta) + c(x))$
- For an iid sample we have $p_\theta(\tilde{x}) = exp(w(\theta)\sum_i T(x_i) - nA(\theta) + c(\tilde{x}))$, where $\sum_i T(x_i) = T(\tilde{x})$ and $nA(\theta) = \tilde{A}(\theta)$.
- If $w(\theta) = \theta$ this is the canonical form
- One can show that $E_\theta[T(\tilde{x})] = \tilde{A}'(\theta)$ and $Var(T(\tilde{x})) = \tilde{A}''(\theta)$ (Proof via moment generating functions).
- The score can be written as $S(\theta) = w'(\theta)T(\tilde{x}) - \tilde{A}'(\theta)$ and the MLE is thus given by
$$\hat{\theta} = \{\theta : w'(\theta)T(\tilde{x}) = \tilde{A}'(\theta)\}$$
- In the canonical case $\hat{\theta} = \{\theta : T(\tilde{x}) = \tilde{A}'(\theta)\}$
- The information is $I(\theta) = -w''(\theta)T(\tilde{x}) + \tilde{A}''(\theta)$. In the canonical case $I(\theta) = \tilde{A}''(\theta)$
- The Fisher information is $FI(\theta) = -w''(\theta)E_\theta[T(\tilde{X})] + \tilde{A}''(\theta)$ which in the canonical case is $\tilde{A}''(\theta)$.
- So, $I(\theta) = FI(\theta)$ in the canonical case.
- Now, we know that $E_\theta[S(\theta)] = 0 = E_\theta[w'(\theta)T(\tilde{X}) - \tilde{A}'(\theta)]$ and so $w'(\theta)E_\theta[T(\tilde{X})] = \tilde{A}'(\theta)$ for all θ.
- We have that $FI(\theta) = \tilde{A}''(\theta) - w''(\theta)E_\theta[T(\tilde{X})]$ where $w'(\theta)E_\theta[T(\tilde{X})] = \tilde{A}'(\theta)$
- On the other hand, at the MLE we have $I(\hat{\theta}) = \tilde{A}''(\hat{\theta}) - w''(\hat{\theta})T(\tilde{x})$, where $w'(\hat{\theta})T(\tilde{x}) = \tilde{A}'(\hat{\theta})$.
- Thus, $I(\hat{\theta}) = FI(\hat{\theta})$

1.1 Multiparameter case

What if $\theta = \{\theta_1, \cdots, \theta_p\}$? Well, the score function $S(\theta)$ is now a vector of derivatives with respect to one parameter at a time.

$$S(\theta) = \begin{pmatrix} \frac{d}{d\theta_1} l(\theta) \\ \cdots \\ \frac{d}{d\theta_p} l(\theta) \end{pmatrix}$$

The observed information is a $p \times p$ matrix with elements $I_{ij}(\theta) = \frac{-d^2}{d\theta_i d\theta_j} l(\theta)$, and the Fisher information contains elements $FI(\theta) E_\theta[\frac{-d^2}{d\theta_i d\theta_j} l(\theta)]$

One can show that $E_\theta[S(\theta)] = 0_{p\times 1}$ and $Var_\theta[S(\theta)] = FI(\theta)$, where FI is the $p \times p$ matrix with elements $FI_{ij}(\theta)$.

A CR-bound can be established in the multiparameter case as well.

CR bound for multiple parameter case: Let $W(\tilde{X})$ be an unbiased estimator of scalar $g(\theta)$ (that is, g is a function that maps the p-dimensional θ into another parameter. We then have that

$$Var_\theta(W(\tilde{X})) \geq \nabla_\theta g(\theta)^T FI(\theta)^{-1} \nabla_\theta g(\theta)$$

In the special case $g(\theta) = (\theta_1, 0, \cdots, 0)$ we have that

$$Var_\theta(\hat{\theta}_1) \geq FI(\theta)^{-1}_{11}$$

It is very important to note that

$$FI(\theta)^{-1}_{11} \neq FI(\theta_1)^{-1}$$

where the RH term denotes the Fisher information for the one-parameter problem. In fact, in general

$$\frac{1}{FI(\theta_1)} \leq FI(\theta)^{-1}_{11}$$

meaning that it is easier to estimate one parameter and estimation precision of θ_1 is affected by estimation of other parameters. In special cases, the FI is diagonal meaning the estimation precision decouples for the different parameters.

2 Intro to asymptotics

We have that $E_{\theta_0}[S(\theta_0)] = 0$ and $Var_{\theta_0}[S(\theta_0)] = FI(\theta_0)$, where I write θ_0 to emphasize that we're integrating w.r.t. the true value of θ.

Now, the score $S(\theta) = \sum_i \log p_\theta(x_i)$ is a r.v., and by the above it has mean 0 when we integrate with respect to the true θ_0 (i.e. match where we evaluate the score and what we average it with respect to). Let us approximate the loglikelihood function $l(\theta)$ near some value θ_0:

$$l(\theta) = l(\theta_0) + (\theta - \theta_0) l'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 l''(\theta_0)$$

and so

$$S(\theta) = S(\theta_0) + (\theta - \theta_0) l''(\theta_0)$$

Specifically, at $\hat{\theta}$

$$0 = S(\hat{\theta}) = S(\theta_0) - (\hat{\theta} - \theta_0) l''(\theta_0)$$

Equivalently, we can write

$$(\hat{\theta} - \theta_0) = \frac{S(\theta_0)}{-l''(\theta_0)} = \frac{S(\theta_0)}{I(\theta_0)},$$

where $I(\theta_0)$ is the observed information at θ_0.
We can rewrite this as

$$(I(\theta_0))^{-1/2}S(\theta_0) = I(\theta_0)^{1/2}(\hat{\theta} - \theta_0)$$

We will now derive the sampling distribution for the RH side of the above equation, which thus gives us the sampling distribution for the LH side, i.e. the sampling distribution of the MLE.

- We have that $S(\theta) = \sum_i \frac{d}{d\theta} logp_\theta(x_i) = \sum_i y_i$ for iid r.v. y_i.
- We know that $E_{\theta_0}[y_i] = 0$ and $E_{\theta_0}[\sum_i y_i] = 0$ under regularity conditions.
- We also know that $Var_{\theta_0}[y_i] = FI(\theta_0)$ where $FI(\theta) = E_\theta[-\frac{d}{d\theta^2} logp_\theta(X)]$ and so $Var_{\theta_0}[\sum_i y_i] = nFI(\theta_0)$

Central Limit Theorem: For iid r.v. Z_i with mean μ and variance σ^2 we define the sum $S_n = \sum_{i=1}^n Z_i$.
Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \rightarrow^d N(0,1)$$

as n goes to infinity.
We will here apply the CLT to $S(\theta)$. That is,

$$(nFI(\theta_0))^{-1/2}S(\theta_0) \rightarrow^d N(0,1)$$

There is just one small problem with the above result. In our approximation of the score, we have $I(\theta_0)$ not $nFI(\theta_0)$. That is, we have the observed information at the MLE which is an *estimate* of the expected information (true variance of the score). Can we just replace $nFI(\theta)$ with $I(\theta_0)$ and expect the CLT still to hold?

Well, we have by the weak law of large numbers (WLLN) that

$$\frac{1}{n}\sum_i \frac{d^2}{d\theta^2} logp_\theta(x_i) \rightarrow E_\theta[\frac{d^2}{d\theta^2} logp_\theta(X)]$$

and so

$$\frac{I(\theta)}{n} \rightarrow^p FI(\theta) \quad for\ all\ \theta$$

Moreover, we have the following theorem:

Slutsky's theorem:

- If $X_n \rightarrow^d X$ (convergence in distribution)
- and $A_n \rightarrow^p a$, where a is a constant. (convergence in probability)
- then $A_n X_n \rightarrow^d aX$

Here we have that

$$(I(\theta_0))^{-1/2}S(\theta_0) = (nFI(\theta_0))^{-1/2}S(\theta_0)(\frac{I(\theta_0)/n}{FI(\theta_0)})^{-1/2}$$

where by the CLT the first term converges to $N(0,1)$ and the second converges in probability to 1 (appealing to the WLLN and the continuous mapping theorem for the function of a r.v. converging in probability), and so by S

Finally, we can thus conclude that

$$(I(\theta_0))^{1/2}(\hat{\theta} - \theta_0) \rightarrow^d N(0,1)$$

Now, this is not an entirely practical result since $I(\theta_0)$ depends on the true value of the parameter. However, relying on the *consistency* (to be shown next) of $\hat{\theta}$) and that $I()$ is continuous in θ, the following result is also true:

$$(I(\hat{\theta}))^{1/2}(\hat{\theta} - \theta_0) \rightarrow^d N(0,1)$$

We can use this to construct confidence intervals for θ as

$$[\hat{\theta} \pm z_{1-\alpha/2} I(\hat{\theta})^{-1/2}]$$

Sometimes it is more convenient to use the following alternative results:

$$\sqrt{nFI(\hat{\theta})}(\hat{\theta} - \theta_0) \rightarrow^d N(0,1)$$

That is, here we replace the observed information by the Fisher information.

We can arrive at the same result if we instead approximate the loglikelihood near $\hat{\theta}$ above, again relying on the consistency of $\hat{\theta}$ is the last part of the proof.
What's the take home message? The sampling distribution of $\hat{\theta}$ is asymptotically normal. Moreover, the estimation variance of $\hat{\theta}$ is well approximated by the inverse of the Fisher information for large n ($Var(\hat{\theta}) = (nFI(\hat{\theta}))^{-1}$) or by the inverse of the observed information. The result also holds in the multiple parameter case when $FI(\theta)$ is a $p \times p$ matrix.

Consistency of the MLE

Consistency means that

$$P(|\hat{\theta} - \theta| > \epsilon) - \rightarrow 0$$

or

$$P(|\hat{\theta} - \theta| \leq \epsilon) - \rightarrow 1$$

That is, eventually $\hat{\theta}$ is in a small neighborhood of the true θ.

This is not a trivial result. The MLE is the maximizer of the observed loglikelihood. To prove that it has the right limit we will have to show that we have equality of the max of a limit and the limit of a max.

We start with some notation:

- Normalized loglikelihood $\bar{l}(\theta|\tilde{X}) = \frac{1}{n}\sum_i logp_\theta(X_i)$ (here written as a random variable)
- By the WLLN we have that $\bar{l}(\theta) \rightarrow^p E_{\theta_0}[l(\theta|X)] = \int (logp_\theta(x))p_{\theta_0}(x)dx$
- We denote this last term by $l_0(\theta)$
- That is, for large n we expect the observed likelihood $\bar{l}(\theta)$ to be close to its expected value $l_0(\theta)$. We therefore expect the maximum of $\bar{l}$ to be close to the maximum of l_0.

What is the maximum of l_0? Here we use the *Information inequality* or the Kullback-Leibler (KL) divergence.

The KL measures the discrepancy between two models f and g, with respect to model f (so KL is not symmetric). That is

$$KL(f,g) = E_f[log\frac{f(X)}{g(X)}] = \int (log\frac{f(x)}{g(x)})f(x)dx$$

Now, we can show that $KL(f,g) \geq 0$ with equality only when $f = g$. The results follows from applying Jensen's inequality.

- Jensen's inequlity. For r.v. Z and convex function $h()$ we have $E[h(Z)] \geq h(E[Z])$
- Here, use $h(z) = -log(z)$
- $KL(f,g) = E_f[-log\frac{g(X)}{f(X)}] \geq -logE_f[\frac{g(X)}{f(X)}] = -log\int g(x)dx = 0$

We now apply this result to the pdf p_θ compared to the true model p_{θ_0} and obtain

$$E_{\theta_0}[-\log\frac{p_\theta(X)}{p_{\theta_0}(X)}] \geq 0$$

or

$$l_0(\theta_0) \geq l_0(\theta)$$

and so, θ_0 is the maximizer of the expected loglikelihood $l_0(\theta)$.

We now need to show that the sequence of estimators $\hat{\theta}(X^n)$ converges to θ_0 in probability (here I use the notation X^n instead of $\tilde{X}$ to emphasize the dependency on the sample size). We already have that

- $\hat{\theta}(X^n)$ is the maximizer of $\bar{l}(\theta|X^n)$
- θ_0 is the maximizer of $l_0(\theta)$
- $\bar{l}(\theta|X^n) \to^p l_0(\theta)$ by WLLN for all θ

See, this is where we need one more step: that the limit of the maximizer ($limit_n\theta(\hat{X}^n)$) is the maximizer of the limit (θ_0). What is needed to make this leap?

- Compactness of the parameter space Θ (a technical condition, trying to guarantee that $\bar{l}$ doesn't act crazy for unbounded θ etc).
- Uniform convergence for $\bar{l}$
$$P_{\theta_0}(\sup_{\theta\in\Theta}|\bar{l}(\theta|X^n) - l_0(\theta)| > \epsilon) \to 0$$
- We also need some conditions on $p_\theta, l(\theta), l_0(\theta)$ continuous in θ and that θ_0 is the unique maximizer of l_0.

Here is a proof:

- We want to prove that $\hat{\theta}(X^n)$ is guaranteed to be in a smaller neighborhood $\Theta(\epsilon) = (\theta_0 - \epsilon, \theta_0 + \epsilon)$ eventually (for large enough n).
- Consequently, we need to make sure that X^n eventually is such that $\bar{l}(\theta)$ is maximized in $\Theta(\epsilon)$
- To prove this we consider two possibilities: $\Theta(\epsilon)$ (what we want) and $\Theta(\epsilon)^c$ (all of Θ not in $\Theta(\epsilon)$)
- As n goes to ∞, can we show that $\hat{\theta}(X^n)$ falls in $\Theta(\epsilon)$ with probability 1?
- Well, this depends on how $\bar{l}(\theta|X^n)$ behaves in both $\Theta(\epsilon)$ and $\Theta(\epsilon)^c$.

The proof is rather cute. We will first find a set of events that are eventually true, i.e. have probability 1 in the limit. Then we will show that if these events are true, then $\hat{\theta}$ has to be in the desired neighborhood.

We start with the region we don't want to end up in: $\Theta(\epsilon)^c$. Now, we assume that l_0 is continuous and that this region of θ is compact, then l_0 attains *some* maximum at *some* value $\theta^* \in \Theta(\epsilon)^c$.
Since θ_0 is the unique maximizer of $l_0(\theta)$ we have that

$$l_0(\theta^*) < l_0(\theta_0)$$

or

$$l_0(\theta_0) - l_0(\theta^*) = \delta > 0$$

for some δ. This is what we know for the *expected* likelihood. Now to the sample outcome:

We define an event

$$A(X^n) = [\sup_{\theta \in \Theta(\epsilon)^c} |\bar{l}(\theta|X^n) - l_0(\theta)| < \delta/2$$

i.e., samples X^n such that the observed loglikelihood and the expected loglikelihood functions are not very far apart for any θ in this region. (Note, the event is defined for the same δ as above).

For such X^n, i.e. when $A(X^n)$ is true, then

$$\bar{l}(\theta) < l_0(\theta) + \delta/2$$

(I just open up the interval inside A and pick on direction). But

$$\bar{l}(\theta) < l_0(\theta) + \delta/2 \geq l_0(\theta^*) + \delta/2 = l_0(\theta_0) - \delta + \delta/2 = l_0(\theta_0) - \delta/2$$

from the results on l_0 above.

What this says is that, if $A(X^n)$ is true, then for $\theta \in \Theta(X^n)$ we have

$$\bar{l}(\theta) < l_0(\theta_0) - \delta/2$$

We now look in the other region: $\Theta(\epsilon)$. Here we know that $l_0(\theta)$ attains its maximum at θ_0. We define a sample event

$$B(X^n) = [\sup_{\theta \in \Theta(\epsilon)} |\bar{l}(\theta|X^n) - l_0(\theta)| < \delta/2]$$

That is, when $B(X^n)$ defines sample events where the observed and expected likelihood are close in the entire range of $\theta \in \Theta(\epsilon)$. Note, $A(X^n)$ and $B(X^n)$ are sample events defined through different values of θ (remember, you can evaluate the likelihood, observed and expected at different θ), so one might be true for a given X^n and one might not.

If $B(X^n)$ is true then

$$\bar{l}(\theta|X^n) > l_0(\theta) - \delta/2$$

(here I just open up the interval inside the B and look in the other direction compared with what I did for A). In particular,

$$\bar{l}(\theta_0|X^n) > l_0(\theta_0) - \delta/2$$

since θ_0 is among the θ where the sample event B is evaluated.

For finite n, we can have $A(X^n)$ and/or $B(X^n)$ be true so we cannot really say that the maximizer of $\bar{l}(\theta)$ will fall in $\Theta(\epsilon)$ with probability 1. However, when we let n go to ∞, the uniform convergence assumption on $\bar{l}$ guarantees that *both* A and B are true with probability 1.

$$P_{\theta_0}(A(X^n) \cup B(X^n)) = P_{\theta_0}(\sup_{\Theta} |\bar{l}(\theta) - l_0(\theta)| \leq \epsilon) \to 1$$

But if both A and B are true, we have

- $\Theta(\epsilon)$: $\bar{l}(\theta) > l_0(\theta) - \delta/2$ and $\bar{l}(\theta_0) > l_0(\theta_0) - \delta/2$
- $\Theta(\epsilon)^c$: $\bar{l}(\theta) < l_0(\theta) - \delta/2$

But we get $\hat{\theta}$ by *maximizing* $\bar{l}(\theta)$ and so, comparing the two bullets above, we will have to have $\hat{\theta}$ in $\Theta(\epsilon)$. That is,

- If A and B are true, $\hat{\theta}$ is in $\Theta(\epsilon)$
- But by the uniform convergence assumption on $\bar{l}$, A and B are true eventually with probability 1
- Therefore $P_{\theta_0}(\hat{\theta} \in \Theta(\epsilon)) \to 1$ and $\hat{\theta}$ is consisten.

Proving the uniform convergence could be a more difficult nut to crack, but for nice pdfs it's not a problem.