Maximum Likelihood

Sofia Olhede

EPFL

October 12, 2020

1. Maximum Likelihood

ML Estimation–Assumptions

- Provided it exists, the MLE of the natural parameter in a k-parameter natural exponential family with open parameter space Φ is consistent.
- Assuming we can get consistency, we can focus on understanding the sampling distribution of the MLE.
- For simplicity, assume $X_1, \ldots, X_n$ are iid with density/frequency) $f(x; \theta)$ for $\theta \in \Theta$. Write
- Let $\ell(x_i; \theta) = \log f(x_i; \theta)$.
- Let $\ell'(x_i; \theta)$, $\ell''(x_i; \theta)$ and $\ell'''(x_i; \theta)$ denote the partial derivatives wrt θ.
- We need some regularity conditions:
- A1: Θ is an open subset of $\mathbb{R}$.
- A2: The support of f, $\operatorname{supp} f$ is independent of θ.
- A3: f is thrice continuously differentiable w.r.t. θ for all the support of f.

ML Estimation–Assumptions II

- A4: $\mathbb{E}(\ell'(X_i;\theta)) = 0$ for all θ and $\mathbb{V}\text{ar}\{\ell'(X_i;\theta)\} = \mathcal{I}_1(\theta)$.
- A5: $-\mathbb{E}(\ell''(X_i;\theta)) = \mathscr{I}_1(\theta)$.
- A6: $\exists M(x) > 0$ and $\delta > 0$ such that $\mathbb{E}\, M(x) < \infty$ and

$$|\theta - \theta_0| < \delta \Rightarrow |\ell'''(x;\theta)| < M(x).$$

ML Estimation–Explaining Assumptions

- If Θ is an open set then for θ_0 the true parameter, it always makes sense for an estimator $\widehat{\theta}_n$ to have a symmetric distribution around θ_0 (such as the Gaussian).
- Under condition (A2) we have

$$\frac{d}{d\theta} \int_{\mathrm{supp}\, f} f(x;\theta)\, dx = 0.$$

This means that we are permitted to exchange integration and differentiation. Therefore it follows

$$0 = \int \frac{d}{d\theta} f(x;\theta) dx = \int \ell'(X_i;\theta) f(x;\theta) = \mathbb{E}\{\ell'(X_i;\theta)\}.$$

Thus once A2 and A4 hold, then we can exchange the order of the limits; and it ensures a finite variance.
- The second derivate has a finite moment by A5.
- A2 and A6 are assumptions that allow us to simplify ("linearize") our understanding of the MLE.

ML Estimation–Explaining Assumptions II

- Taking our exchange of limits even further: if we can differentiate twice then

$$0 = \int \frac{d}{d\theta}\left\{\ell'(x;\theta) f(x;\theta)\right\} dx$$
$$= \int \ell''(x;\theta) f(x;\theta)\, dx + \int \left\{\ell'(x;\theta)\right\}^2 f(x;\theta)\, dx.$$

Thus we may deduce that $\mathcal{I}(\theta) = \mathscr{I}(\theta)$

ML Estimation–Distribution

- Theorem: Let $X_1, \ldots, X_n$ be IID random variables with the same density $f(x;\theta)$. Assume that A1–A6 are satisfied. If the MLE $\widehat{\theta}_n$ exists and is unique, and we have consistency then

$$\sqrt{n}\left\{\widehat{\theta}_n - \theta\right\} \xrightarrow{\mathcal{L}} N\left(0, \mathcal{I}_1(\theta)/\mathscr{I}_1^2(\theta)\right).$$

  Furthermore, when we can say that $\mathcal{I}(\theta) = \mathscr{I}(\theta)$ then

$$\sqrt{n}\left\{\widehat{\theta}_n - \theta\right\} \xrightarrow{\mathcal{L}} N(0, 1/\mathcal{I}_1(\theta)).$$

- For finite samples we often say

$$\widehat{\theta}_n \overset{d}{\approx} N(\theta, 1/\mathcal{I}_n(\theta)).$$

- Thus for large enough samples, the MLE is approximate Gaussian, approximately unbiased, and approximately achieving the Cramér–Rao lower bound.

ML Estimation–Distribution

- Proof: Assuming (A1)–(A3) we can note that

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

- We now implement a Taylor series to deduce that

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

$$+ \frac{1}{2}(\hat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*). \tag{2}$$

We have terminated the Taylor series after 3 terms. To make the equality hold, that means we need an exact form for the remainder. We here use the Lagrane form, other forms include Cauchy. θ_n^* lies between θ and $\widehat{\theta}_n$.

ML Estimation–Distribution

- Dividing the last equation by $\sqrt{n}$ means we obtain

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(X_i;\theta) + (\hat{\theta}_n - \theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell''(X_i;\hat{\theta}_n) + \frac{1}{2\sqrt{n}}(\hat{\theta}_n - \theta)^2\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*). \tag{3}$$

- Now, we use this equation to re-write

$$(\hat{\theta}_n - \theta) = -\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(X_i;\theta)}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell''(X_i;\hat{\theta}_n) + \frac{1}{2\sqrt{n}}(\hat{\theta}_n - \theta)\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*)}.$$

- We can now use the Central Limit Theorem (CLT) that we just derived to deduce

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(X_i;\theta) \xrightarrow{d} N(0, 1/\mathcal{I}_n(\theta)).$$

ML Estimation–Distribution

- Furthermore, we can note that (from the Law of Large Numbers):

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell''(X_i; \hat{\theta}_n) \xrightarrow{p} -\mathscr{I}(\theta).$$

- We define the remainder term to be

$$R_n = \frac{1}{2\sqrt{n}}(\hat{\theta}_n - \theta) \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*).$$

  If we can show that $R_n \xrightarrow{p} 0$ then the denominator tends to $\mathscr{I}(\theta)$ and we can use Slutsky's Lemma to deduce the result.

- We have already showed $\hat{\theta}_n - \theta \xrightarrow{p} 0$ and $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*)$ does not diverge and so we may deduce that $R_n \xrightarrow{p} 0$. Thus the result follows.

ML Estimation–Distribution

- Can we deduce from this result that maximum likelihood estimators are optimal?
- Well this is an asymptotic result, e.g. it holds **eventually** in n.
- For a fixed value of n things are less clear.
- Also we have assumed that we are comparing unbiased estimators- what can be gained by relaxing this assumption?

Shrinkage Estimation

- As simple results by Charles Stein show, things are not as simple as they seem.
- Let $Y_1, \dots, Y_n$ be independent random variables.
- Assume that each $Y_i \sim N(\mu_i, \sigma^2)$. Each Y_i has a different mean, but the variance is coupled.
- First we might consider the slightly simpler case of $\sigma^2 = 1$.
- Then we wish to estimate $\boldsymbol{\mu}$.
- We use mean square error to assess performance.
- This is not quite a "standard" problem as the number of parameters is growing with the sample size.

Shrinkage Estimation II

- The log-likelihood can easily be written up as

$$\ell(\boldsymbol{\mu}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu_i)^2.$$

- Using differentiation we can easily deduce that

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{Y}.$$

- This is like having n MLE's each of sample size 1. Not great estimation; too few samples.
- The MSE is then n as

$$\mathrm{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbb{E}\,\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = n.$$

- Is this the best we can do?

Shrinkage Estimation III

- Assume that $\boldsymbol{Y} = \left(Y_1, \ldots Y_n\right)^T$ such that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \mathrm{I}_n)$ where $\boldsymbol{\mu} \in \mathbb{R}^n$. We call this "Stein's set-up".
- We define a shrinkage estimator to be

$$\widetilde{\boldsymbol{\mu}}_a = \left(1 - \frac{a}{\|\boldsymbol{Y}\|^2}\right)\boldsymbol{Y} = \left(1 - \frac{a}{\|\hat{\boldsymbol{\mu}}\|^2}\right)\hat{\boldsymbol{\mu}}.$$

This is the shrunken version of the MLE. Assume that $n \geq 3$ then

- for all $a \in (0, 2n-4)$

$$\mathrm{MSE}(\widetilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) \leq \mathrm{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}),$$

- for $a = n-2$

$$\mathrm{MSE}(\widetilde{\boldsymbol{\mu}}_{n-2}, 0) \leq \mathrm{MSE}(\widehat{\boldsymbol{\mu}}, 0),$$

- For all $\boldsymbol{\mu} \in \mathbb{R}^n$ and all $a \in (0, 2n-4)$

$$\mathrm{MSE}(\widetilde{\boldsymbol{\mu}}_{n-2}, \boldsymbol{\mu}) \leq \mathrm{MSE}(\widetilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}).$$

Shrinkage Estimation IV

- This was a very surprising result at the time!
- First, the MLE is outperformed.
- The Stein estimator takes the MLE and “shrinks” its magnitude.
- The amount of shrinkage depends on $\|\boldsymbol{Y}\|$.
- This estimate takes the estimate of μ_j into account when estimating μ_i.
- There are no “smoothness” assumptions.
- The performance of the MLE deteriorates as n increases.
- To understand this result we need to go in stages.

Shrinkage Estimation V

- Lemma: let $Y \sim N(\theta, \sigma^2)$ and assume $h: \mathbb{R} \to \mathbb{R}$ is differentiable. If
  1. $\mathbb{E}\,|h(Y)| < \infty$;
  2. $\lim_{y\to\pm\infty}\left\{h(y)\exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)\right\} = 0$ then
     $\mathbb{E}\{h(Y)(Y-\theta)\} = \sigma^2\,\mathbb{E}(h'(Y))$.
- Proof: we note that from the definition of expectation

$$\begin{aligned}
\mathbb{E}\{h(Y)(Y-\theta)\} &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}} h(y)(y-\theta)e^{-\frac{1}{2\sigma^2}(y-\theta)^2} \\
&= \left[-\frac{\sigma}{\sqrt{2\pi}}h(y)e^{-\frac{1}{2\sigma^2}(y-\theta)^2}\right]_{y\to-\infty}^{\infty} \\
&+ \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h'(y)e^{-\frac{1}{2\sigma^2}(y-\theta)^2}\,dy \\
&= \sigma^2\,\mathbb{E}(h'(Y)). \qquad (4)
\end{aligned}$$

This completes the proof of the lemma.

Shrinkage Estimation VI

- Now we are ready to prove the Stein theorem.
- We note that from first principles we have that

$$\begin{aligned}\mathrm{MSE}(\widetilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) &= \mathbb{E}\, \left\| \left(1 - \frac{a}{\|\boldsymbol{Y}\|^2}\right) \boldsymbol{Y} - \boldsymbol{\mu} \right\|^2 \\ &= \mathbb{E}\, \left\| \boldsymbol{Y} - \boldsymbol{\mu} - \frac{a}{\|\boldsymbol{Y}\|^2} \boldsymbol{Y} \right\|^2 \\ &= \mathbb{E}\, \|\boldsymbol{Y} - \boldsymbol{\mu}\|^2 - 2\,\mathbb{E}\left(\frac{a\boldsymbol{Y}^T(\boldsymbol{Y} - \boldsymbol{\mu})}{\|\boldsymbol{Y}\|^2}\right) + \mathbb{E}\left(\frac{a\|\boldsymbol{Y}\|^2}{\|\boldsymbol{Y}\|^4}\right) \\ &= n - 2a \sum_{i=1}^{n} \mathbb{E}\left[\frac{Y_i(Y_i - \mu_i)}{\sum_{j=1}^{n} Y_j^2}\right] + a^2\,\mathbb{E}\left(\frac{1}{\|\boldsymbol{Y}\|^2}\right). \qquad (5)\end{aligned}$$

- To understand this quantity we need to study the middle term.

Shrinkage Estimation VII

- We define n differentiable functions $h_i : \mathbb{R}^n \to \mathbb{R}$ by

$$\boldsymbol{u} = (u_1, \ldots, u_n) \overset{h_i}{\mapsto} \frac{u_i}{u_i^2 + \sum_{j \neq i} u_j^2}.$$

- We observe that for all $i \in \{1, \ldots, n\}$ and all $\{u_j\}_{j \neq i} \in \mathbb{R}^{n-1}$ then

$$\lim_{u_i \to \pm\infty} \left\{ h_i(\boldsymbol{u}) \exp\left[-\frac{1}{2\sigma^2}(u_i - \mu_i)^2 \right] \right\} = 0.$$

- We note that h_i becomes an $\mathbb{R} \to \mathbb{R}$ fn once $\{u_j\}_{j \neq i} \in \mathbb{R}^{n-1}$ is fixed.
- We now use the tower property and apply our lemma to re-write:

$$\mathbb{E}\left[\frac{Y_i(Y_i - \mu_i)}{\sum_{j=1}^n Y_j^2} \right].$$

Shrinkage Estimation VIII

- Applying our lemma we get

$$
\begin{aligned}
\mathbb{E}\left[\frac{Y_i(Y_i-\mu_i)}{\sum_{j=1}^n Y_j^2}\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{Y_i(Y_i-\mu_i)}{\sum_{j=1}^n Y_j^2}|\{Y_j\}_{j\neq i}\right]\right\} \\
&= \mathbb{E}\{\mathbb{E}[h_i(\boldsymbol{Y})(Y_i-\mu_i)|\{Y_j\}_{j\neq i}]\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\left.\frac{\partial}{\partial u_i}h_i(\boldsymbol{u})\right|_{\boldsymbol{u}=\boldsymbol{Y}}\left|\{Y_j\}_{j\neq i}\right.\right]\right\} \\
&= \mathbb{E}\left[\left.\frac{\partial}{\partial u_i}h_i(\boldsymbol{u})\right|_{\boldsymbol{u}=\boldsymbol{Y}}\right] \\
&= \mathbb{E}\left[\frac{\|\boldsymbol{Y}\|^2-2Y_i^2}{\|\boldsymbol{Y}\|^4}\right].
\end{aligned}
\tag{6}
$$

Shrinkage Estimation IX

- It follows that we can re-write the mean square error as:

$$\begin{aligned}\mathrm{MSE}(\widetilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) &= n - 2a\,\mathbb{E}\left[\frac{n\|\boldsymbol{Y}\|^2 - 2\|\boldsymbol{Y}\|^2}{\|\boldsymbol{Y}\|^4}\right] + a^2\,\mathbb{E}\left[\frac{1}{\|\boldsymbol{Y}\|^2}\right] \\ &= n + \left(a^2 - 2a(n-2)\right)\mathbb{E}\left[\frac{1}{\|\boldsymbol{Y}\|^2}\right]. \qquad (7)\end{aligned}$$

- The polynomial $p(a) = a^2 - 2a(n-2)$ is strictly negative in the range $a \in (0, 2n-4)$. This gives part 1.
- On the same range in a $p(a)$ has a unique minimum at $a = n - 2$. This proves part 3.
- For part (2) we note that if $\boldsymbol{\mu} = 0$ then $\|\boldsymbol{Y}\|^2 \sim \chi^2_n$.
- Thus we note that $\mathbb{E}\{\frac{1}{\|\boldsymbol{Y}\|^2}\} = 1/(n-2)$ and recall that $n \geq 3$.
- Therefore it follows $\mathrm{MSE}(\widetilde{\boldsymbol{\mu}}_{n-2}, 0) = 2$.

Loss functions I

- We can also replace the mean square error by another convex measure of performance.
- Thus we replace $\|\widehat{\theta} - \theta\|$ by a choice deviation measure $\mathcal{L}(\widehat{\theta}, \theta)$ called a loss function.
- The expected loss is the <u>the risk</u>:

$$R(\widehat{\theta}, \theta) = \mathbb{E}\{\mathcal{L}(\widehat{\theta}, \theta)\}.$$

- Selecting the right loss function is crucial and this choice must be made carefully.

Loss functions II

- Example: the exponential distribution. Assume that $Y_1, \ldots Y_n \sim \mathrm{Exponential}(\lambda)$ and that $n \geq 2$.
- The MLE of λ is
$$\widehat{\lambda} = \frac{1}{\overline{Y}}.$$
- Here $\overline{Y}$ is the empirical mean.
- We can calculate
$$\mathbb{E}\{\widehat{\lambda}\} = \frac{n\lambda}{n-1}.$$
- It therefore follows that $\widetilde{\lambda} = (n-1)\widehat{\lambda}/n$ is an unbiased estimator. Observe that
$$\mathrm{MSE}(\widetilde{\lambda}) < \mathrm{MSE}(\widehat{\lambda}).$$
Thus $\widehat{\lambda}$ is dominated by $\widetilde{\lambda}$.

Loss functions III

- λ here takes a positive value.
- In this case quadratic estimation penalises over estimation more heavily than underestimation.
- The maximum possible under estimation is bounded.
- But could we change the loss function?
- Instead we could use

$$\mathcal{L}(a, b) = a/b - 1 - \log(a/b).$$

- Note that for all fixed a $\lim_{b \to 0} \mathcal{L}(a, b) = \lim_{b \to \infty} \mathcal{L}(a, b) = \infty$.
- Now for $n > 1$ we calculate the risk

$$R(\lambda, \widetilde{\lambda}) = \mathbb{E}_{\lambda}\left[\frac{n\lambda\bar{Y}}{n-1} - 1 - \log\left(\frac{n\lambda\bar{Y}}{n-1}\right)\right]. \tag{8}$$

Loss functions IV

- The risk is

$$R(\lambda, \widetilde{\lambda}) = \mathbb{E}_\lambda\left[\frac{n\lambda\bar{Y}}{n-1} - 1 - \log\left(\frac{n\lambda\bar{Y}}{n-1}\right)\right]$$
$$= \mathbb{E}_\lambda\left[\lambda\bar{Y} - 1 - \log(\lambda\bar{Y})\right] + \frac{\mathbb{E}_\lambda\left[\lambda\bar{Y}\right]}{n-1} - \log\left(\frac{n}{n-1}\right)$$
$$= \mathbb{E}_\lambda\left[\lambda\bar{Y} - 1 - \log(\lambda\bar{Y})\right] + g(n). \tag{9}$$

- To derive the simplification we write $\bar{Y} = \frac{n-1}{n}\bar{Y} + \frac{1}{n}\bar{Y}$.
- Note that $\mathbb{E}_\lambda(\bar{Y}) = \lambda^{-1}$. Thus

$$g(n) = \frac{1}{n-1} - \log\left(\frac{n}{n-1}\right).$$

- We claim that $g(n) > 0$ once $n \geq 2$.

Loss functions V

- Using that $\log(x) = \int_1^x t^{-1}\, dt$ this follows if

$$\frac{1}{x} > \log(x+1) - \log(x), \quad x > 1$$
$$\Leftrightarrow \frac{1}{x} > \int_x^{x+1} t^{-1} dt, \quad x > 1. \tag{10}$$

This inequality holds by a rectangle area bound on the integral, as follows:

$$\frac{1}{x} = [(1+x) - x]\frac{1}{x} = \int_x^{x+1} \frac{1}{x} dt > \int_x^{x+1} \frac{1}{t} dt,$$

when $x > 1$.

- It therefore follows $R(\widetilde{\lambda}, \lambda) > R(\widehat{\lambda}, \lambda)$ and so $\widetilde{\lambda}$ dominates $\widehat{\lambda}$.

Loss functions VI

- We can push generality even further, and obtain an all encompassing framework.
- Called decision theory, it views inference as a game between nature and the statistician.
- Recall our general framework for statistical inference:
  1. Model phenomenon by distribution $F(y_1, \ldots, y_n; \theta)$ for some $n \geq 1$.
  2. Distributional form is known but $\theta \in \Theta$ is not known.
  3. Observe realisation of $(Y_1, \ldots, Y_n)$ from this distribution.
  4. Use $(Y_1, \ldots, Y_n)$ in order to make assertions concerning the true value of , and quantify the uncertainty associated with these assertions.
- The decision theory framework formalises step (4) to include estimation, testing, and confidence intervals.