Statistics for Data Science: Week 4

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

The first sort of assertion we wish to make is:

1. **Point Estimation**. Given realisation $(Y_1, \ldots, Y_n)^\top$ from $F(y_1, ..., y_n; \theta)$, how can we produce an educated guess for the unknown true parameter θ?

How? With a point estimator!

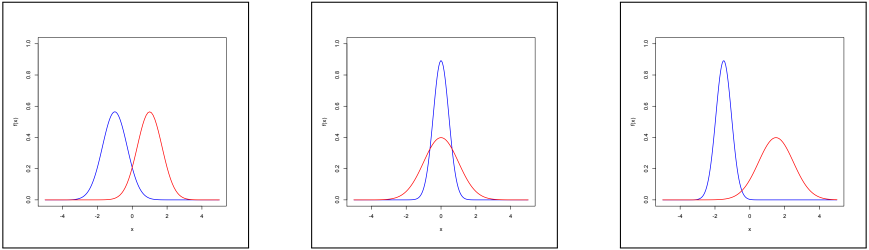
Definition (Point Estimator)

A statistic with codomain Θ is called a *point estimator*, i.e. a point estimator is a statistic $T : \mathcal{Y}^n \to \Theta$.

Since the objective of an estimator is to estimate the θ that generated the data, we typically denote it by $\hat{\theta}(Y_1, ..., Y_n)$, or just $\hat{\theta}$. Note that θ is a deterministic parameter, whereas $\hat{\theta}$ is a random variable.

But which estimator?

- Any statistic taking values in Θ could be used!
- Simpler yet, if we are given some $\hat{\theta}$, how do we judge its quality?
- Since estimators are *random variables*, every different realisation of the sample $(Y_1, \ldots, Y_n)$ will produce a different realised value for $\hat{\theta}$.
- A good estimator should be such that it typically manifests realisations that fall near the true θ.
- More precisely, the sampling distribution of an estimator should be concentrated around the true parameter value θ.



Definition (Mean Squared Error)

Let $\hat{\theta}$ be an estimator of a parameter θ corresponding to a model $\{F_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$. The mean squared error of $\hat{\theta}$ is defined as

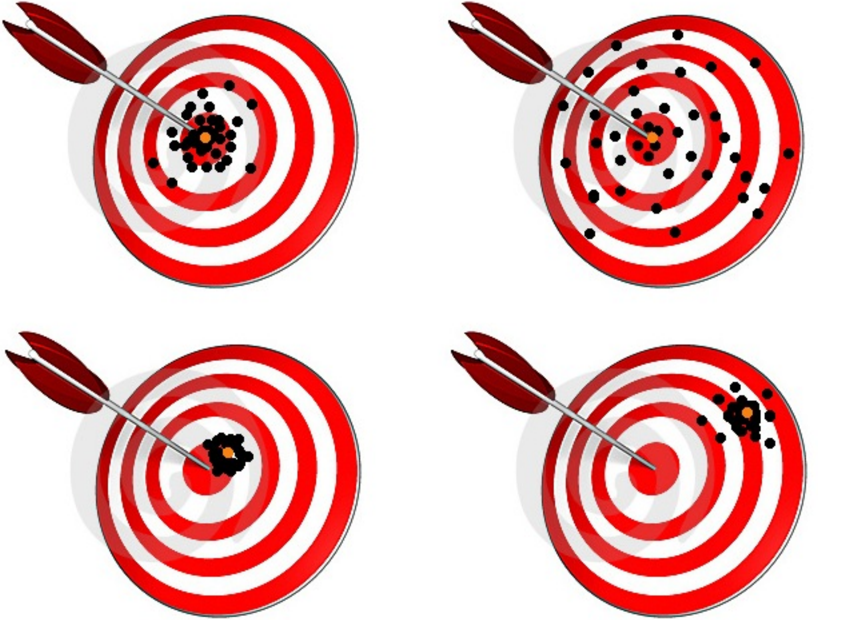
$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right].$$

And here's the relation to means and variances:

Lemma (Bias-Variance Decomposition)

The MSE admits the decomposition

$$\mathrm{MSE}(\hat{\theta}, \theta) = \underbrace{\left\|\mathbb{E}[\hat{\theta}] - \theta\right\|^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\|\hat{\theta} - \mathbb{E}(\hat{\theta})\|^2\right]}_{\text{variance}}.$$



Proof.

We expand the MSE after adding and subtracting $\mathbb{E}[\hat{\theta}]$:

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}-\theta\|^2] &= \mathbb{E}[\|\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta\|^2] \\
&= \mathbb{E}\Big[(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)^\top(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)\Big] \\
&= \|\mathbb{E}[\hat{\theta}]-\theta\|^2+\mathbb{E}\big[\|\hat{\theta}-\mathbb{E}(\hat{\theta})\|^2\big]+2\mathbb{E}\Big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^\top(\mathbb{E}[\hat{\theta}]-\theta)\Big] \\
&= \|\mathbb{E}[\hat{\theta}]-\theta\|^2+\mathbb{E}[\|\hat{\theta}-\mathbb{E}(\hat{\theta})\|^2]+2\underbrace{(\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}])}_{=0}{}^\top(\mathbb{E}[\hat{\theta}]-\theta)
\end{aligned}$$

by linearity of the expectation and since $(\mathbb{E}[\hat{\theta}]-\theta)$ is deterministic. □

As foretold, the concentration of an estimator $\hat{\theta}$ around the true parameter θ can always be bounded by the MSE:

Lemma

Let $\hat{\theta}$ be an estimator of $\theta \in \mathbb{R}^p$. For any $\epsilon > 0$,

$$\mathbb{P}[\|\hat{\theta} - \theta\| > \epsilon] \leq \frac{\mathrm{MSE}(\hat{\theta}, \theta)}{\epsilon^2}$$

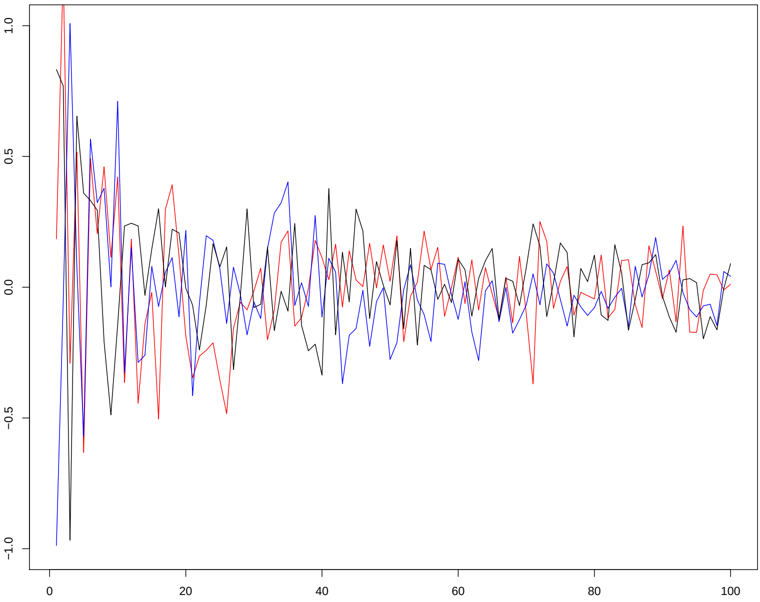
- Note that $\mathrm{MSE}(\hat{\theta}_n, \theta) \overset{n\to\infty}{\longrightarrow} 0 \implies \hat{\theta}_n \overset{p}{\longrightarrow} \theta$.
- When an estimator has this property, we call it consistent.

Definition (Consistency)

An estimator $\hat{\theta}_n$ of θ, constructed on the basis of a sample of size n, is consistent if $\hat{\theta}_n \overset{p}{\longrightarrow} \theta$ as $n \to \infty$.

Note that a vanishing MSE implies consistency, but the converse generally fails.

Consistency of sample mean of sample mean of $Y_1, ..., Y_n \sim \mathcal{N}(0, 1)$, towards the true parameter value 0 (by law of large numbers).



Is it always possible to get consistent estimators?

Depends on whether the estimation problem is well-posed

Definition (Identifiability)

A probability model $\{F_\theta\}_{\theta \in \Theta}$ is called identifiable if for any pair $\theta_1, \theta_2 \in \Theta$ we have the implication

$$\theta_1 \neq \theta_2 \implies F_{\theta_1} \neq F_{\theta_2}.$$

- Lack of identifiability means that the same model can be produced by more than one parameter.
- In this case we could never distinguish amongst the parameters that give the same model.
- Example: if we have $N(\mu_1 + \mu_2, \sigma^2)$, we can never identify each μ_i, but only their sum.
- Henceforth we will tacitly assume identifiability (and make special mention if it is at stake).

- We can use the MSE to compare estimators or to gauge their performance.
- But is there a *best possible MSE* for a given problem?
- This is a very difficult problem, equivalent to finding a uniformly optimal estimator: a statistic T_* such that

$$\mathrm{MSE}(T_*, \theta) \leq \mathrm{MSE}(T, \theta)$$

  for all $\theta \in \Theta$ and all other estimators T.
  - To see this, let $T = c$ be a trivial (constant) estimator and observe that for any non-trivial estimator S we have $\mathrm{MSE}(S, \theta) > \mathrm{MSE}(T, \theta)$ at $\theta = c$.
  - So if we want to do well for all θ we can't do perfectly for any specific θ.
- Here's a simpler question to ask instead (ruling out trivial estimators):

Among unbiased estimators (bias zero), can we make the MSE arbitrarily small?

- If so, how? (what is the crucial ingredient at play?)

- Rephrasing, we are asking whether there is *fundamental lower bound* for the variance of an unbiased estimator of θ.
- We will concetrate on 1-dimensional parameters for simplicity.

The Question

For $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top$ with joint density/frequncy $f(\boldsymbol{y}; \theta)$ depending on an unknown $\theta \in \mathbb{R}$, does there exist some function $\Lambda(\theta) > 0$ such that

$$\mathrm{var}[\hat{\theta}] \geq \Lambda(\theta), \quad \forall \theta$$

for any estimator $\hat{\theta}$ such that $\mathbb{E}[\hat{\theta}] = \theta$?

- At a next step we can ask if this bound is achievable.

Let's assume we can interchange differentiation and integration in the form

$$\frac{d}{d\theta}\int S(\mathbf{y})f(\mathbf{y};\theta)d\mathbf{y} \overset{!}{=} \int S(\mathbf{y})\frac{f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\frac{d}{d\theta}f(\mathbf{y};\theta)d\mathbf{y} = \int S(\mathbf{y})f(\mathbf{y};\theta)\frac{d}{d\theta}\log f(\mathbf{y};\theta)d\mathbf{y}$$

whenever an integral such as the one on the left hand side presents itself.

1. Setting $U = \frac{\partial}{\partial\theta}\log f(\mathbf{Y};\theta)$ and $S(\mathbf{y}) = 1$, this gives that

$$\mathbb{E}[U] = \int f(\mathbf{y};\theta)\frac{\partial}{\partial\theta}\log f(\mathbf{y};\theta)d\mathbf{y} = \frac{d}{d\theta}\int f(\mathbf{y};\theta)d\mathbf{y} = 0$$

2. Therefore $\mathrm{var}[U] = \mathbb{E}[U^2] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{Y};\theta)\right)^2\right]$

3. For $\hat{\theta}$ unbiased, our "interchange ansatz" with $S(\mathbf{y}) = \hat{\theta}(\mathbf{y})$ gives

$$\mathrm{cov}\left(\hat{\theta}, U\right) = \mathbb{E}[\hat{\theta}\,U] - \mathbb{E}[\hat{\theta}]\underbrace{\mathbb{E}[U]}_{=0} = \int \hat{\theta}(\mathbf{y})f(\mathbf{y};\theta)\frac{d}{d\theta}\log f(\mathbf{y};\theta)d\mathbf{y} = \frac{d}{d\theta}\underbrace{\mathbb{E}[\hat{\theta}]}_{=\theta} = 1$$

Now the Cauchy-Schwartz inequality gives

$$\mathrm{var}(\hat{\theta}) \geq \frac{\mathrm{cov}^2\left(\hat{\theta}, U\right)}{\mathrm{var}\left(U\right)} \implies \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{Y};\theta)\right)^2\right]}$$

In summary, we have established:

Cramér-Rao Lower Bound

Given sufficient regularity, any unbiased estimator $\hat{\theta}(\boldsymbol{Y})$ of finite variance satisfies:

$$\mathrm{var}[\hat{\theta}(\boldsymbol{Y})] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right)^2\right]} = \frac{1}{\mathcal{I}_n(\theta)}$$

The quantity $\mathcal{I}_n(\theta)$ is fundamental, and called Fisher information.

- If $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ has iid entries, we have $f(\boldsymbol{y};\theta) = \prod_{i=1}^n f(y_i;\theta)$ and so

$$\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta).$$

- By further interchanges of integration/differentiation it typically holds that

$$\mathcal{I}_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{Y};\theta)\right]$$

- The deeper meaning of all this will become clearer when we study likelihood.

Is the Cramér-Rao bound tight (achievable) though?

$$\text{if} \quad \mathrm{var}[\hat{\theta}] = \frac{1}{\mathcal{I}_n(\theta)}$$

$$\text{then} \quad \mathrm{var}[\hat{\theta}] = \frac{\mathrm{cov}^2\left[\hat{\theta}, \frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right]}{\mathrm{var}\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right]}$$

which occurs if and only if $\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)$ is a linear function of $\hat{\theta}$ (correlation 1):

$$\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta) = A(\theta)\hat{\theta}(\boldsymbol{Y}) + B(\theta)$$

Solving this differential equation yields, for all $\boldsymbol{y}$,

$$\log f(\boldsymbol{y};\theta) = A^*(\hat{\theta}) + B^*(\theta) + S(\boldsymbol{y})$$

so that $\mathrm{var}_\theta(\hat{\theta})$ attains the lower bound if and only if the density (frequency) of $\boldsymbol{Y}$ is a one-parameter exponential family with sufficient statistic $\hat{\theta}$.

So what ingredients go into pushing towards this lower bound?

The Rao-Blackwell Theorem tells us that sufficiency is key:

Theorem (Rao-Blackwell Theorem)

Let $\hat{\theta}$ be an unbiased estimator of θ with finite variance, and let T be sufficient for θ. Then $\hat{\theta}^* := \mathbb{E}[\hat{\theta}|T]$ is also an unbiased estimator of θ and

$$\mathrm{var}(\hat{\theta}^*) \leq \mathrm{var}(\hat{\theta}).$$

Equality is attained if and only if $\mathbb{P}_\theta[\hat{\theta}^* = \hat{\theta}] = 1$.

Comments:

- Throwing away irrelevant aspects of the data improves estimation quality.
- These irrelevant aspects contribute to the variation of the estimator (as they have sampling variation of their own), but without furnishing any useful information on the parameter
- $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|T]$ is called a "Rao-Blackwellised" version of $\hat{\theta}$.

Proof.

Since T is sufficient for θ, $\mathbb{E}[\hat{\theta}|T = t] = h(t)$ is independent of θ, so that $\hat{\theta}^*$ is well-defined as a statistic (depends only on $\boldsymbol{Y}$ and not θ). Then,

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\mathbb{E}[\hat{\theta}|T]] = \mathbb{E}[\hat{\theta}] = \theta.$$

Furthermore, from the law of total variance, we have

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}[\mathbb{E}(\hat{\theta}|T)] + \mathbb{E}[\mathrm{var}(\hat{\theta}|T)] \geq \mathrm{var}[\mathbb{E}(\hat{\theta}|T)] = \mathrm{var}(\hat{\theta}^*)$$

In addition, note that

$$\mathrm{var}(\hat{\theta}|T) := \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}|T])^2|T] = \mathbb{E}[(\hat{\theta} - \hat{\theta}^*)^2|T]$$

so that $\mathbb{E}[\mathrm{var}(\hat{\theta}|T)] = \mathbb{E}(\hat{\theta} - \hat{\theta}^*)^2 > 0$ unless if $\mathbb{P}(\hat{\theta}^* = \hat{\theta}) = 1$. □

Suppose that $\hat{\theta}$ is an unbiased estimator of $g(\theta)$ and T, S are θ-sufficient.

- What is the relationship between $\underbrace{\mathrm{var}(\mathbb{E}[\hat{\theta}|T])}_{\hat{\theta}^*_T} \overset{?}{\gtreqless} \underbrace{\mathrm{var}(\mathbb{E}[\hat{\theta}|S])}_{\hat{\theta}^*_S}$
- Intuition suggests that whichever of T, S carries the least irrelevant information (in addition to the relevant information) should "win"
  - ↪ More formally, if $T = h(S)$ then we should expect that $\hat{\theta}^*_T$ dominate $\hat{\theta}^*_S$.

Proposition

For $\hat{\theta}$ an unbiased estimator of θ and T, S two θ-sufficient statistics, define

$$\hat{\theta}^*_T := \mathbb{E}[\hat{\theta}|T] \quad \& \quad \hat{\theta}^*_S := \mathbb{E}[\hat{\theta}|S].$$

Then, the following implication holds

$$T = h(S) \implies \mathrm{var}(\hat{\theta}^*_T) \leq \mathrm{var}(\hat{\theta}^*_S)$$

- Essentially this means that the best possible "Rao-Blackwellisation" is achieved by conditioning on a minimally sufficient statistic.

Proof.

Recall the *tower property* of conditional expectation: if $Y = f(X)$, then

$$\mathbb{E}[Z|Y] = \mathbb{E}\{\mathbb{E}(Z|X)|Y\}.$$

Since $T = f(S)$ we have

$$\begin{aligned} \hat{\theta}^*_T &= \mathbb{E}[\hat{\theta}|T] \\ &= \mathbb{E}[\mathbb{E}(\hat{\theta}|S)|T] \\ &= \mathbb{E}[\hat{\theta}^*_S|T] \end{aligned}$$

The conclusion now follows from the Rao-Blackwell theorem. □

- So now we have a means to judge the quality of an estimator
- In certain cases, we even know what's the best performance we can hope for.
- And (minimal) sufficiency can help us approach it.
- But how can we actually come up with an estimator in the first place?
- We need general methods that can be applied in any model context to yield an estimator.
- Preferably methods that yield good estimators relative to our performance measures/bounds.

$\hookrightarrow$ The main focus will be on a key method called maximum likelihood.

Motivation: recall our understanding of statistics as “inverse probability”.

$\hookrightarrow$ For the moment, consider the discrete case for simplicity.

Probability Perspective

Given a parameter $\theta \in \Theta$, then for any $(y_1, ..., y_n)^\top \in \mathcal{Y}^n$, we can evaluate

$$(y_1, ..., y_n) \mapsto \mathbb{P}_\theta[Y_1 = y_1, ..., Y_n = y_n]$$

that is, how the probability varies as a function of the sample (=the result).

Statistics Perspective

Given a sample $(y_1, ..., y_n)^\top \in \mathcal{Y}^n$, then for any $\theta \in \Theta$ we can calculate

$$\theta \mapsto \mathbb{P}_\theta[Y_1 = y_1, ..., Y_n = y_n]$$

that is, how the probability varies as a function of θ (=the model).

Intuition: we imagine that, having our sample, the values of θ that are most plausible are those that render the observed sample most probable...

This motivates the following definition...

Definition (Likelihood)

Let $(Y_1, \ldots, Y_n)$ be a sample of random variables with joint density/frequency $f(y_1, ..., y_n; \theta)$, where $\theta \in \mathbb{R}^p$. The likelihood of θ is defined as

$$L(\theta) = f(Y_1, ..., Y_n; \theta).$$

If $(Y_1, ..., Y_n)^\top$ has i.i.d. entries, each with density/frequency $f(y_i; \theta)$ then,

$$L(\theta) = \prod_{i=1}^{n} f(Y_i; \theta)$$

... and the following estimation method

Definition (Maximum Likelihood Estimator)

In the same context, a maximum likelihood estimator (MLE) of $\hat{\theta}$ is an estimator such that

$$L(\theta) \leq L(\hat{\theta}), \qquad \forall\, \theta \in \Theta.$$

Many comments are in order:

- When there exists a unique maximum, we speak of the MLE $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$
- The likelihood is a random function.
- It is the joint density/frequency of the sample, but viewed as a function of θ.
- It is NOT "the probability of θ"
- $L(\theta)$ is the answer to the question *how does the joint density/probability of the sample vary as we vary θ?*
  - In the discrete case it is exactly "the probability of observing our sample" as a function of θ.
  - In the continuous case, since $F(\boldsymbol{y} + \varepsilon/2\,; \theta) - F(\boldsymbol{y} - \varepsilon/2; \theta) \approx \|\varepsilon\| f(\boldsymbol{y}; \theta)$ as $\|\varepsilon\| \downarrow 0$, we can view $\|\epsilon\| \times L(\theta)$ as being the "probability of observing something in the neighbourhood of our sample", as a function of θ.

- If a sufficient statistic T exists for θ then Fisher-Neyman factorisation implies

$$L(\theta) = g(T(\boldsymbol{Y}); \theta)h(\boldsymbol{Y}) \propto g(T(\boldsymbol{Y}); \theta)$$

  i.e. any MLE depends on data only through a sufficient statistic.

- Since the sufficient statistic was arbitrary, if a minimally sufficient statistic exists, the MLE will have used an estimator that has achieved the maximal sufficient reduction of the data.

- MLE's are also *equivariant*. If $g : \Theta \to \Theta'$ is a bijection, and if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ (you can take the hat out: $g(\hat{\theta}) = \widehat{g(\theta)}$)

- When the likelihood is differentiable in θ, its maximum $L(\theta)$ must solve the equation

$$\nabla_\theta L(\theta) = 0,$$

- But before declaring a solution as an MLE, we must verify it to be a maximum (and not a minimum!).

- If the likelihood is twice differentiable in θ, we can verify this by checking

$$-\nabla^2_\theta L(\theta)\big|_{\theta=\hat{\theta}} \succ 0,$$

i.e that minus the Hessian is positive definite. In one dimension, this reduces to the standard second derivative criterion.

- To solve $\nabla_\theta L(\theta) = 0$ when the Y_i are independent, we must painfully calculate the derivative of an n-fold product.

- To avoid this, we focus instead on the loglikelihood $\ell(\theta) := \log L(\theta)$ instead. Maximisation of ℓ is equivalent to maximisation of L by monotonicity.

- When the Y_i are independent, ℓ has the advantage of being a sum rather than a product

$$\ell(\theta) = \log\left(\prod_{i=1}^{n} f_{Y_i}(Y_i;\theta)\right) = \sum_{i=1}^{n} \log f_{Y_i}(Y_i;\theta).$$

- Of course, under twice differentiability, verification of a maximum can be checked again by whether or not

$$\nabla_\theta \ell(\theta)|_{\theta=\hat{\theta}} = 0 \quad \& \quad -\nabla^2_\theta \ell(\theta)\big|_{\theta=\hat{\theta}} \succ 0.$$

Example (MLE for Bernoulli trials)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(p)$. The likelihood is

$$L(p) = \prod_{i=1}^{n} f(Y_i; p) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} = p^{\sum_{i=1}^{n} Y_i}(1-p)^{n-\sum_{i=1}^{n} Y_i}$$

giving loglikelihood

$$\ell(p) = \log p \sum_{i=1}^{n} Y_i + \log(1-p)\left(n - \sum_{i=1}^{n} Y_i\right).$$

This is twice differentiable in p and we calculate

$$\frac{d}{dp}\ell(p) = p^{-1}\sum_{i=1}^{n} Y_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} Y_i\right).$$

Example (MLE for Bernoulli trials, continued)

Solving

$$p^{-1}\sum_{i=1}^{n} Y_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} Y_i\right) = 0,$$

we get the unique root $\frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$. Calling this $\hat{p}$, we now verify that

$$\frac{d^2}{dp^2}\ell(p) = -p^2\sum_{i=1}^{n} Y_i - (1-p)^{-2}\left(n - \sum_{i=1}^{n} Y_i\right),$$

which is a negative expression, since $0 \leq \sum_{i=1}^{n} Y_i \leq n$ andt $p \in (0,1)$. Thus

$$\hat{p} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is the unique MLE of p. □

Example (MLE for exponential distribution)

Let $Y_1, \dots, Y_n \overset{iid}{\sim} Exp(\lambda)$. The likelihood is

$$L(\lambda) = \prod_{i=1}^{n} f(Y_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda Y_i} = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} Y_i\right\}.$$

and the log likelihood is

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} Y_i.$$

This is twice differentiable in λ and we calculate

$$\frac{d}{d\lambda}\ell(\lambda) = n\lambda^{-1} - \sum_{i=1}^{n} Y_i.$$

Example (MLE for exponential distribution, continued)

Setting $\ell'(\lambda) = 0$ we get a unique root

$$\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^{-1} = 1/\bar{Y}.$$

Call this $\hat{\lambda}$, and note that

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{n}{\lambda^2}$$

is always negative, since $\lambda > 0$. Thus

$$\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^{-1} = 1/\bar{Y}$$

is the unique MLE of λ. □

Example (MLE for Gaussian distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(Y_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{2\sigma^2}\right\}.$$

giving loglikelihood

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

All partial second derivatives exist and are

$$\begin{aligned} \frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu) \\ \frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \mu)^2. \end{aligned}$$

Example (MLE for Gaussian distribution, continued)

Solving $\nabla_{(\mu,\sigma^2)}\ell(\mu,\sigma^2) = 0$ for (μ,σ^2) gives a system of equations in two unknowns, with unique root

$$\left(\bar{Y}, n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right).$$

Call this $(\hat{\mu},\hat{\sigma}^2)$, and let's verify it's a maximum. Note that

$$\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2) = -\frac{n}{\sigma^2}, \quad \frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(Y_i - \mu)^2$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2) = \frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2) = -\frac{\sum_{i=1}^{n}(Y_i - \mu)}{\sigma^4} = \frac{n\mu - n\bar{Y}}{\sigma^4}.$$

Calculating these derivatives at $(\hat{\mu},\hat{\sigma}^2)$, we get

$$\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = -\frac{n}{\hat{\sigma}^2}, \quad \frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = -\frac{n}{2\hat{\sigma}^4}$$

Example (MLE for Gaussian distribution, continued)

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = \frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = \frac{n\hat{\mu}-n\hat{\mu}}{\hat{\sigma}^4} = 0.$$

Thus the matrix

$$\left[-\nabla^2_{(\mu,\sigma^2)}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}\right]$$

is diagonal. If both of its diagonal elements are positive, then it will be positive definite. This is indeed the case since $\hat{\sigma}^2 > 0$ nd so the unique MLE of (μ,σ^2) is given by

$$(\hat{\mu},\hat{\sigma}^2) = \left(\bar{Y}\ ,\ \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right).$$

□

Note that from our Gaussian sampling results we get that σ^2 is biased.

Example (MLE for Poisson Distribution)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Then

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{Y_i}}{Y_i!} e^{-\lambda} \right\} \implies \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \log(Y_i!)$$

Setting $\nabla_\theta \log L(\theta) = -n + \lambda^{-1} \sum Y_i = 0$ we obtain $\hat{\lambda} = \bar{Y}$ since $\nabla_\theta^2 \log L(\theta) = -\lambda^{-2} \sum Y_i < 0$.

Example (MLE for Uniform Distribution – a non-differentiable case)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. The likelihood is

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq Y_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq Y_{(n)}\}.$$

Hence if $\theta < Y_{(n)}$ the likelihood is zero. In the domain $[Y_{(n)}, \infty)$, the likelihood is a decreasing function of θ. Hence $\hat{\theta} = Y_{(n)}$.

Example (Equivariance of the MLE)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$, and suppose we're interested in estimating $\mathbb{P}[Y_1 \leq y]$, for a given $y \in \mathbb{R}$. Note that

$$\mathbb{P}[Y_1 \leq y] = \mathbb{P}[Y_1 - \mu \leq y - \mu] = \Phi(y - \mu),$$

where Φ is the standard normal CDF. The mapping $\mu \mapsto \Phi(y - \mu)$ is bijective, since Φ is strictly monotone. So by equivariance, the MLE of $\mathbb{P}[Y_1 \leq y]$ is $\Phi(y - \hat{\mu})$, where $\hat{\mu}$ is the MLE of μ (which by our previous example is $\hat{\mu} = \bar{Y}$).

Example (Equivariance and usual vs natural parameterisation)

Let $Y_1, \ldots, Y_n \stackrel{iid}{\sim} f$, with

$$f(y) = \exp\{\phi T(y) - \gamma(\phi) + S(y)\}, \qquad y \in \mathcal{Y}$$

where $\phi \in \Phi \subseteq \mathbb{R}$ is the natural parameter. Suppose we can write $\phi = \eta(\theta)$, where $\theta \in \Theta$ is the usual parameter and $\eta : \Theta \to \Phi$ is a differentiable bijection (so that $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, for $d = \gamma \circ \eta$). In this notation, the density/frequency takes the form

$$\exp\{\phi T(y) - \gamma(\phi) + S(y)\} = \exp\{\eta(\theta) T(y) - d(\theta) + S(y)\}.$$

Equivariance now implies that if $\hat{\theta}$ is the MLE of θ, then $\eta(\hat{\theta})$ is the MLE of $\phi = \eta(\theta)$. The converse is also true: if $\hat{\phi}$ is the MLE of ϕ, then $\eta^{-1}(\hat{\phi})$ is the MLE of $\theta = \eta^{-1}(\phi)$. □

Examples show that likelihood generally gives sensible estimators – still:

- Beyond intuition, is there a canonical mathematical reason for it?
- What rigorous guarantees can we offer?
  - $\hookrightarrow$ Can we get consistency?
  - $\hookrightarrow$ Can we approach reasonable MSE performance?

To answer these questions, we go back to entropy and Kullback-Leibler divergence.