Maximum Likelihood Review

Maximal likelihood estimation is a method of estimating the parameters of a statistical model given data.

By statistical model on a sample space $\mathcal{X}$ we mean a set of distributions (actually measures) on $\mathcal{X}$. If we write $PM(\mathcal{X})$ for the space of all possible distributions (measures) over $\mathcal{X}$ then a model is a subset $M \subset PM(\mathcal{X})$. The elements of M are indexed by a parameter θ with values in a parameter space T, that is,

$$M = \{P_\theta \mid \theta \in T\}$$

where each P_θ is a member of the set $PM(\mathcal{X})$.

A model is parametric if T is finite dimensional. Usually $T \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$. If $\dim(T) = \infty$ then M is a non-parametric model.

The canonical problem of statistics is to take observations $x_1, \ldots, x_N$, $x_n \in \mathcal{X}$, which we model as random variables $X_1, \ldots X_N$ which we assume are drawn from P_θ, i.e.

$$X_1, \ldots, X_N \sim_{iid} P_\theta \qquad \theta \in T$$

and use them to tell us something about the value of θ.

Estimators

Assume there is a sample $X_1, X_2 \ldots X_n$ of n iid observations coming from a true but unknown distribution $P_0(\cdot)$.

Let us assume that $P_0 \in M$, ie $\exists\ \theta_0 \in T$ s.t. $P_{\theta_0} = P_0$. Let us also assume that T is finite dimensional.

We would like an estimator $\hat{\theta} = f(X_1, \ldots, X_N)$ which is close to θ_0.

An estimator is a statistic, ie. a function of the data, that is used to infer the value of an unknown parameter in a statistical model. Some estimators can be quite complex.

Let's name some important prop's of estim's: Letting $X = \{x_1, \ldots x_n\}$ and $\hat{\theta} = f(X)$ assumed 1-dimensional

1) $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2]$

- note that this expectation is with respect to samples (populations of size n) drawn from P_0

this is the mean or expected squared error; small if the estimator is good

2) $\text{var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2]$

this is, for a given sample size (and P_0) what is the characteristic "spread" of the estimand

Datasets / samples

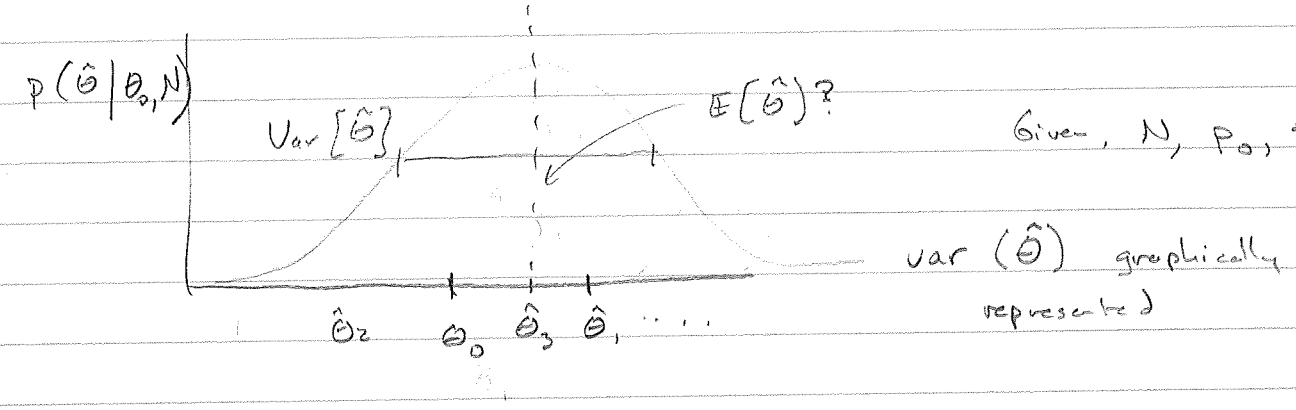
$X_1 = \{x_1 = \cdot \quad x_2 = \cdot\cdot \quad x_3 = \cdot\cdot \quad x_N = \cdots\}$, Estimands $\hat{\theta}_1 =$

$X_2 =$ $\hat{\theta}_2$

$X_3 :$ $\hat{\theta}_3$

$\vdots$ $\vdots$

with $x_n \sim iid\ P_\theta$, $P_\theta(X) = \prod P_\theta(x)$



3) Bias $B(\hat{\theta}) = E(\hat{\theta}) - \theta_0$

is the distance between the average of estimates over all samples of a given size and θ_0.

Statisticians like unbiased estimators,

ie. $B(\hat{\theta}) = 0$

Note for now the relationship:

$$MSE(\hat{\theta}) = var(\hat{\theta}) + (Bias(\hat{\theta}))^2$$

proof

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta_0)^2] \\ &= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}) - \theta_0)^2] \\ &= E[(\hat{\theta} - E[\hat{\theta}])^2] \\ &\quad (A) + E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta_0)] \\ &\quad (B) + E[(E[\hat{\theta}] - \theta_0))(\hat{\theta} - E[\hat{\theta}])] \\ &\quad + E[(E[\hat{\theta}] - \theta_0)^2] \end{aligned}$$

notice that $E[\hat{\theta}]$, θ_0 are constant

means that $(A) = (E[\hat{\theta}] - E[\hat{\theta}])(E[\hat{\theta}] - \theta_0) = 0$

and same for (B). So

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta}] - \theta_0)^2 \\ &= VAR(\hat{\theta}) + (BIAS(\hat{\theta}))^2 \end{aligned}$$

We will return to this later.

Let's look at a simple estimator of a population mean and variance (think heights if you like). Assume the true distribution is $N(\mu_0, \sigma_0^2)$. Let N be the sample size and $X = \{x_1, \dots x_N\}$, $x_n \sim iid\ N(\mu, \sigma^2)$

Let's consider some estimators for μ

A) $\hat{\mu}_A = f_A(x_1, \dots x_N) = 0$

B) $\hat{\mu}_B = f_B(x_1, \dots, x_N)$, $\hat{\mu}_B = \underset{\mu}{\text{argmax}}\ \mathcal{L}(X;\mu)$

Maximum Likelihood →

C) $\hat{\mu}_C = f_C(x_1 \dots x_N) = \lambda \mu_1 + (1-\lambda)\hat{\mu}_B$

↑ some "guess" $\mu_1 \approx \mu_0$

A) $\text{Var}(\hat{\mu}_A) = 0$

$\text{BIAS}(\hat{\mu}_A) = \hat{\mu}_0$

$\text{MSE}(\hat{\mu}_A) = \hat{\mu}_0^2$

B) This is the maximum likelihood estimator where we have a parametric model

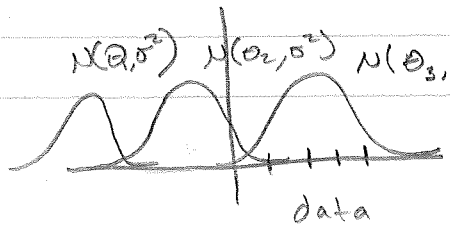
$$x \sim N(\mu, \sigma^2), \quad \Theta = \{\mu, \sigma^2\}$$

and

$$\mathcal{L}(X;\Theta) = \prod_{n=1}^{N} p(x_n|\Theta)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

Likelihood function →



which Θ is preferred?

ML Estimation by easy example

If we want to extremize $\mathcal{L}(x;\theta)$ we look for

$$\frac{\partial \mathcal{L}(x;\theta)}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 \mathcal{L}(x;\theta)}{\partial \theta^2} \leq 0$$

to maximize $\mathcal{L}(x;\theta)$ w.r.t. θ

Important $\frac{\partial}{\partial \theta}\mathcal{L}(x;\theta)$ is almost always nasty so it is usual to work with $\frac{\partial}{\partial \theta}\log \mathcal{L}(x;\theta)$ which has the same max because log is monotonically increasing. Often the resulting maximization is easier

Claim

$$\arg\max_\theta \log(\mathcal{L}(\theta)) = \arg\max_\theta (\mathcal{L}(\theta))$$

Check 0 points

$$0 = \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \frac{1}{\mathcal{L}(\theta)} \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \quad \cup \quad \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = 0 \Leftrightarrow \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$$

Check curvature

$$\operatorname{sign}\left(\frac{\partial^2}{\partial \theta^2}\log \mathcal{L}(\theta)\right) = \operatorname{sign}\left(\frac{\partial}{\partial \theta}\frac{1}{\mathcal{L}(\theta)}\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right)$$

$$= \operatorname{sign}\left(-\frac{1}{\mathcal{L}(\theta)^2}\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\frac{\partial \mathcal{L}(\theta)}{\partial \theta} + \frac{1}{\mathcal{L}(\theta)}\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta}\right)$$

$$= \operatorname{sign}\left(\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\left(\underbrace{\frac{1}{\mathcal{L}(\theta)} - \frac{1}{\mathcal{L}(\theta)^2}}_{\text{true if pos.}}\right)\right)$$

$$\frac{1}{\mathcal{L}(\theta)} - \frac{1}{\mathcal{L}(\theta)^2} \stackrel{?}{\geq} 0$$

$$\mathcal{L}(\theta)^2 - \mathcal{L}(\theta) \geq 0 \quad \mathcal{L}(\theta)^2 \geq \mathcal{L}(\theta) \;\cup$$

Back to mean estimator, B

$$\arg\max \log h(x;\theta)$$

occurs at $\frac{\partial}{\partial\theta}\log h(x;\theta) = 0$ here $\theta = \mu$

$$\frac{\partial}{\partial\mu}\log h(x;\mu) = \sum_n \frac{\partial}{\partial\mu}\log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$= \sum_n \frac{\partial}{\partial\mu} -\frac{(x_n-\mu)^2}{2\sigma^2}$$

$$= \sum_n 2(x_n-\mu)$$

$$= 2\left(\sum_n x_n - n\mu\right) = 0$$

$$\Rightarrow \hat{\mu}_B = \frac{\sum_{n=1}^{N} x_n}{N}$$

Bias of estimator B:

$$E[\hat{\mu}_B] = E\left[\frac{\sum_{n=1}^{N} x_n}{N}\right] = \frac{1}{N}\sum_{n=1}^{N} E[x_n]$$

$$= \frac{1}{N}\cdot N\cdot\mu = \mu$$

$\hat{\mu}_B$ is unbiased, ie. $\text{BIAS}(\mu_B) = 0$

Variance of estimator:

$$\text{Var}[\hat{\mu}_B] = \text{Var}\left[\frac{\sum_{n=1}^{N} x_n}{N}\right] = \frac{1}{N^2}\text{Var}\left[\sum_{n=1}^{N} x_n\right]$$

$$= \frac{1}{N^2}\sum_{n=1}^{N}\text{Var}(x_n)$$

$$= \frac{\sigma^2}{N}$$

So $\text{MSE}(\hat{\mu}_B) = \text{VAR}(\hat{\mu}_B) = \frac{\sigma^2}{N}$ which, in the limit of $N\to\infty$, $\frac{\sigma^2}{N}\to 0$.

Estimation : the core of stats, along with frequentist inference.

Assuming that we know σ_0^2 let's hypothesize that $x_n \sim_{iid} N(\phi, \sigma_0^2)$ then to test this hypothesis we run the following thought experiment. Assume that we use estimator B and that we draw an infinite number of sample populations of size N from the model above.

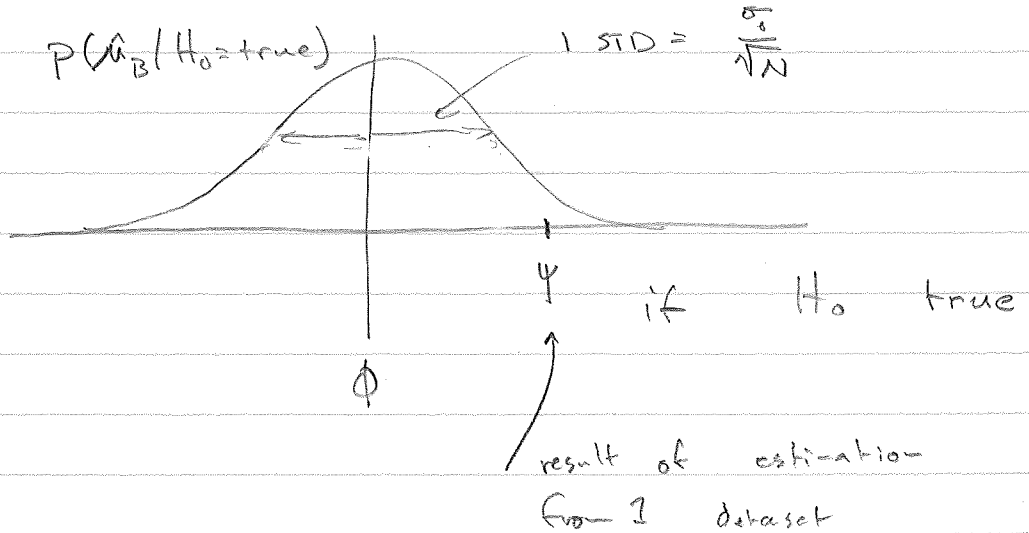
We already know that, in this case,

$$E[\hat{\mu}_B] = \phi \quad \text{and} \quad Var[\hat{\mu}_B] = \frac{\sigma_0^2}{N}$$

What can be demonstrated is that, for sufficiently large N and under mild regularity assumptions

$$\hat{\mu}_B \sim N\left(\phi, \frac{\sigma_0^2}{N}\right)$$

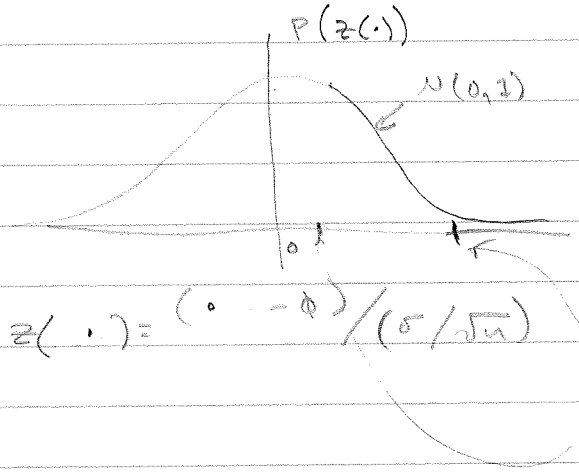
Now, given a particular dataset X_D we compute $\hat{\mu}_B(X_D) = \psi$. We have a picture like this



To think about confidence intervals and drawing conclusions about a null hypothesis it will help, in this and many cases to "normalize" our estimator to a z-score, ie. a $N(0,1)$ variable.

If $\hat{\mu}_B \sim N(\phi, \sigma^2/n)$ then

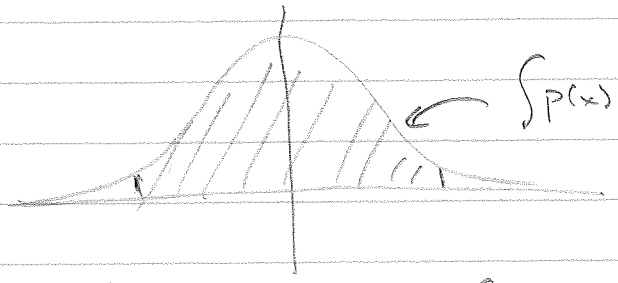
$$z(\cdot) = \frac{(\hat{\mu}_B - \phi)}{\sigma/\sqrt{n}} \sim N(0,1)$$ Pf. trivial



Frequentist inference rejects a null hypothesis when the value of an estimator computed on a real sample is surprising (i.e. low probability) under the null hypothesis.

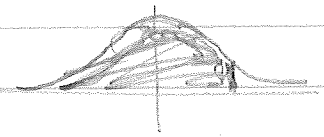
If, for instance, $z(\psi)$ falls within the 95% confidence interval then we fail to reject H_0. What is the 95% confidence interval?

It's, under the standard normal



i.e. the value z for which
$p(-z \le Z \le z) = 1-\alpha = .95$ for $Z \sim N(0,1)$

This occurs at $\Phi(z) = P(Z \le z) = 1 - \frac{\alpha}{2} = 0.975$
where $\Phi(z)$ is the CDF of $N(0,1)$



$$\Phi(z) = \int_0^z N(x;0,1)\,dx$$

i.e. $\Phi^{-1}(\Phi(z) = 0.975) = 1.96$ so

$$0.95 = 1-\alpha = P(-z \le Z \le z) = P\left(-1.96 \le \frac{\Psi - \phi}{\sigma/\sqrt{N}} \le 1.96\right)$$

and, thusly

$$0.95 = P\left(\Psi - 1.96\cdot\left(\frac{\sigma}{\sqrt{N}}\right) \le \phi \le \Psi + 1.96\left(\frac{\sigma}{\sqrt{N}}\right)\right)$$

What does this mean? It means that both ϕ, under our assumptions, is likely to be in this band but, perhaps more importantly, that ϕ and Ψ can be flipped saying, in effect, that if the true parameter is ϕ, Ψ outside of

$$\left(\phi - 1.96\left(\frac{\sigma}{\sqrt{N}}\right),\ \phi + 1.96\left(\frac{\sigma}{\sqrt{N}}\right)\right)$$

would be surprising and could be taken

as evidence (at 5% c.f.) to reject the null hypothesis.

Note that a p-value is the value of α one would have to choose to reject the null hypothesis given the observed sample, i.e. p-value is soln to (in this case)

$$\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \frac{\Psi-\phi}{\sigma/\sqrt{n}} \quad \text{or} \quad 1-\frac{\alpha}{2} = \Phi\left(\frac{\Psi-\phi}{\sigma/\sqrt{n}}\right)$$

$$\text{p-value} = \alpha = 2\left(1-\Phi\left(\frac{\Psi-\phi}{\sigma/\sqrt{n}}\right)\right)$$

Reminding all the way back we have estimators

	A	B	C
var	0	$\frac{\sigma^2}{n}$	
bias	$\hat{\mu}_A$	0	
mse	$\hat{\mu}_A^2$	$\frac{\sigma^2}{n}$	

Which suggests than a) no one estimator is always best and b) that we might be able to explore the spectrum of estimators and engineer estimators that work well for our problem.

For instance, what if we think we know the true "answer", is. $\mu_1 \approx \mu_0$, like estimator C.

$$\lambda\mu_1 + (1-\lambda)\hat{\mu}_B$$

Estimator C

$$\text{Bias}\left(\lambda\mu_0 + (1-\lambda)\hat{\mu}_B\right)$$

$$= E\left[\lambda\mu + (1-\lambda)\hat{\mu}_B\right] - \mu_0$$

$$= \lambda\mu_1 + (1-\lambda)\mu_0 - \mu_0 = \lambda(\mu_1 - \mu_0)$$

$$\text{Var}\left(\lambda\mu_1 + (1-\lambda)\hat{\mu}_B\right)$$

$$= (1-\lambda)^2 \text{Var}(\hat{\mu}_B) = (1-\lambda)^2 \frac{\sigma^2}{N}$$

	A	B	C
var	0	σ^2/N	$(1-\lambda)^2 \frac{\sigma^2}{N}$
bias	$\hat{\mu}_A$	0	$\lambda(\mu_1 - \mu_0)$
mse	$\hat{\mu}_A^2$	σ^2/N	$(1-\lambda)^2 \frac{\sigma^2}{N} + \lambda^2(\mu_1-\mu_0)^2$

But, wait, this suggests that bias could lead to lower MSE, and here will when

$$\sigma^2/N > (1-\lambda)^2 \frac{\sigma^2}{N} + \lambda^2(\mu_1-\mu_0)^2$$

$$\sigma^2/N\left(1-(1-\lambda)^2\right) > \lambda^2(\mu_1-\mu_0)^2$$

$$\sigma^2/N > \frac{\lambda^2}{2\lambda-\lambda^2}(\mu_1-\mu_0)^2$$

$$\sigma^2/N > \frac{1}{\frac{2}{\lambda}-1}(\mu_1-\mu_0)^2$$

$$\sigma^2/N > \frac{\lambda}{2-\lambda}(\mu_1-\mu_0)^2$$

which could easily happen.

The take-home here is that maximum likelihood estimators are but one family of estimator and, further, that bias is not a bad word, in fact introducing it can lower the variance of your estimator and potentially the MSE of your estimator as well.

Frequentist inference involves computing and using confidence intervals for the sample variance of estimators.

Bias introduced in the form of regularization and priors may make for better estimation provided that the introduced bias is helpful.