Model selection

- So far, assumed we were working with the true model family $\mathcal{P}$ → estimation problem just one of fixing parameter values.

What happens if we use the wrong model!

Return to Information inequality

Assume the model is g, and we use model f.

(Example g is t-distribution and f is normal)

Measure of model distance KL

$$KL(g,f) = E_g \log \frac{g(x)}{f(x)} = \int \left(\log \frac{g(x)}{f(x)}\right) g(x)\, dx \geq 0$$

(shown before)

with equality if $g = f$

$$\Rightarrow \quad E_g \log g(X) \geq E_g \log f(X)$$

so expected log likelihood is greater under the model than wrong model (wrt truth)

What happens if we proceed with MLE using f?

$$L_f(\theta|X^n) = \prod_{i=1}^{n} f_\theta(x_i) \Rightarrow \hat{\theta} = \text{argmax} \underbrace{\frac{1}{n}\sum_{i=1} \log f_\theta(x_i)}$$

$$\xrightarrow[\text{WLLN}]{} E_g[\log f_\theta(X)]$$

note g - since that's the data generating process

$$\text{So} \max_\theta L_f(\theta|X^n) \overset{\text{large } n}{=} \min_\theta \left(-E_g[\log f_\theta(X)]\right)$$

$$= \min_\theta \underbrace{E_g \log g(X)} - E_g[\log f_\theta(X)]$$

unknown constant

$$= \min_\theta KL(g, f)$$

$\Rightarrow$ so MLE equivalent to finding closest f to g among $\{f_\theta : \theta \in \Theta\}$.

As $n \to \infty$ $\hat{\theta} = \text{argmax}_\theta L_f(\theta|X^n) \Rightarrow$ converges to $\text{argmax}_\theta E_g[\log f_\theta(X)]$

$= \theta^*$

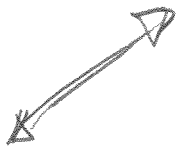
Where f_{θ^*} closest model to the model g.

Note, θ^* may have nothing to do with g itself.

When parameter θ, and estimate $\hat{\theta} \to \theta^*$ where θ^* is meaningful i.e. θ^* corresponds to some population parameter, like mean & variance, under g as well, the modelling is called <u>robust</u>.

(Example from book $g \sim \text{Gamma}(4,1)$ mean $4 \cdot 1 = 4$
Variance $4 \cdot 1^2 = 4$

$f \sim N(\theta, \sigma^2)$



turns out $\hat{\theta} \to 4$
$\hat{\sigma}^2 \to 4$)

But no free lunch $\Rightarrow$ tends to be the case that $\hat{\theta}_f$ less efficient than $\hat{\theta}_g$.

<u>Consistency result here</u>

$\theta_0 = \text{argmax}\ E_g \log f(X)$

Under reg. conditions $\hat{\theta} = \text{argmax}\ \bar{l}_f(\theta | X^n)$

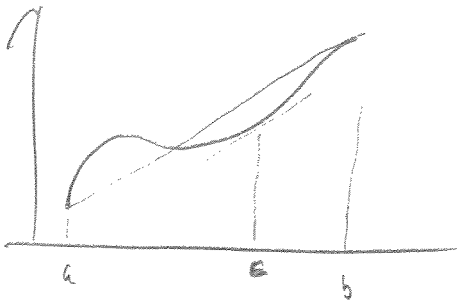
$= \text{argmax}\ \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i) \xrightarrow{p} \theta_0$

(Similar proof as before)

What about other asymptotic results?

- $S_f(\theta) = \text{Score} = \frac{\partial}{\partial\theta} \log L_f(\theta|X^n) = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \log f_\theta(x_i)$
- $\hat{\theta}_f = \text{argmax}\ S_f(\theta)$
- Expand Score around $\hat{\theta} \Rightarrow S_f(\theta) \approx S_f(\hat{\theta}) + \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\hat{\theta}} (\theta - \hat{\theta})$

Mean-value theorem



$$\frac{f(b)-f(a)}{b-a} = f'(c)$$

some $c \in [a,b]$

$\Rightarrow$ Here, take θ^* somewhere between θ and $\hat{\theta}$

$$\Rightarrow S_f(\theta) = S_f(\hat{\theta}) + \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta^*} (\theta - \hat{\theta})$$

$$= \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta^*} (\theta - \hat{\theta})$$

- Now $S_f(\theta) = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \underbrace{\log f_\theta(x_i)}_{y_i}$, where $E(Y_i) = E_g\left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right]$

$$= \frac{\partial}{\partial\theta} E_g[\log f_\theta(X)] = \frac{\partial}{\partial\theta} \int (\log f_\theta(x)) g(x)\, dx$$

for nice g & f

and since θ_0 = max. maximizer of $E_g \log f \Rightarrow E(Y_i) = 0$ @ $\theta = \theta_0$

Moreover $\mathrm{Var}(Y_i) = E\left[\left(\frac{\partial \log f_\theta}{\partial \theta}\right)\left(\frac{\partial \log f_\theta}{\partial \theta}\right)'\right]\Big|_{\theta=\theta_0} = \mathcal{J}(\theta_0)$
@ $\theta=\theta_0$

and by CLT $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \to^d N(0, \mathcal{J})$

Now return to score

$$S_f(\theta) = \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta^*} (\theta - \hat{\theta})$$

$$n \cdot \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(X_i)\Big|_{\theta^*}\right)$$

$$\to^p E\left(\frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(X)\right)\Big|_{\theta_0} = -\mathcal{I}(\theta_0)$$

since $\hat{\theta} \to \theta_0$
and θ^* between $\hat{\theta}$ and θ_0

$$\Rightarrow \frac{1}{\sqrt{n}} S_f(\theta_0) = -\frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta^*} \cdot \sqrt{n}(\hat{\theta} - \theta_0)$$

$\to^d N(0, \mathcal{J}(\theta_0))$ $\qquad$ $\to^p \mathcal{I}(\theta_0)$

$$\Rightarrow \sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \mathcal{I}^{-1}(\theta_0)\mathcal{J}(\theta_0)\mathcal{I}^{-1}(\theta_0))$$

$= \mathcal{I}^{-1}(\theta_0)$ if $f_{\theta_0} = g$

So, asymptotic variance is $I^{-1}(\theta_0) J(\theta_0) I^{-1}(\theta_0)$

note, function of true unknown θ_0 and unknown g!

so of course not to practical to use

$\Rightarrow$ but can examine for model classes to get some idea about (loss of) efficiency

Breakdown of ML Principle

General idea was that maximizing likelihood $\Rightarrow$ θ "most likely" to generate data observed

↳ "Bayesian view"

BUT, now assume we explore a class of models $\{f \in \mathcal{F}\}$ with parameter dimension $\{p_f, f \in \mathcal{F}\}$

Simple case: nested models $f_1 \subset f_2 \subset f_3 \ldots \in f_M$

e.g. regression models w. β_1, (β_1, β_2), $(\beta_1, \beta_2, \beta_3)$ ...