Fisher information

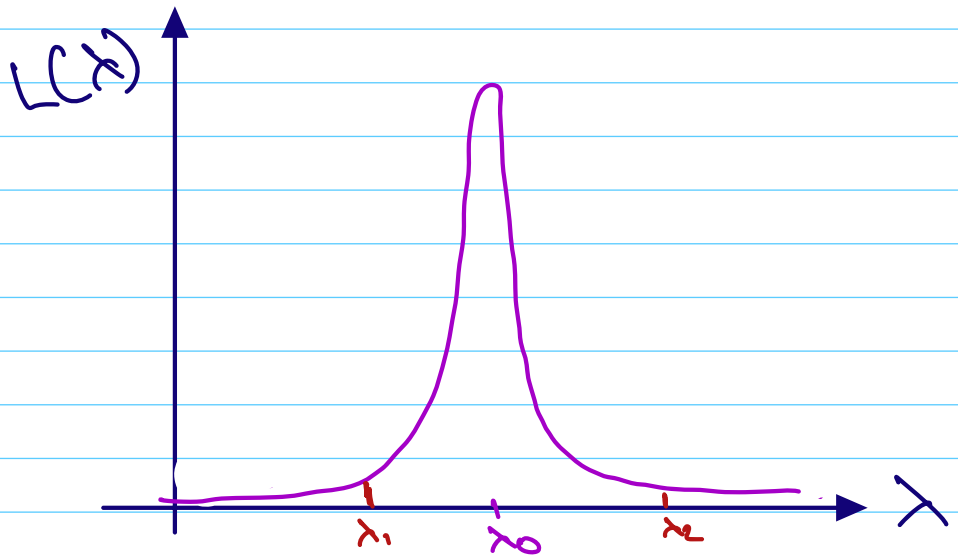
likelihood function is the basis of a lot of modern statistics and science and its defined as follows.

$$L(\theta \mid Data) \equiv \mathbb{P}(Data \mid \theta) = \mathbb{P}(Data \mid theory)$$

So we are asking having made a measurement how likely is it that we got measurement, Given that the universe is described by certain theory (θ)

For example:

$$L(\lambda \mid x_1, x_2, x_3 .. x_n) = \mathbb{P}(x_1, x_2 .. x_n \mid \lambda)$$

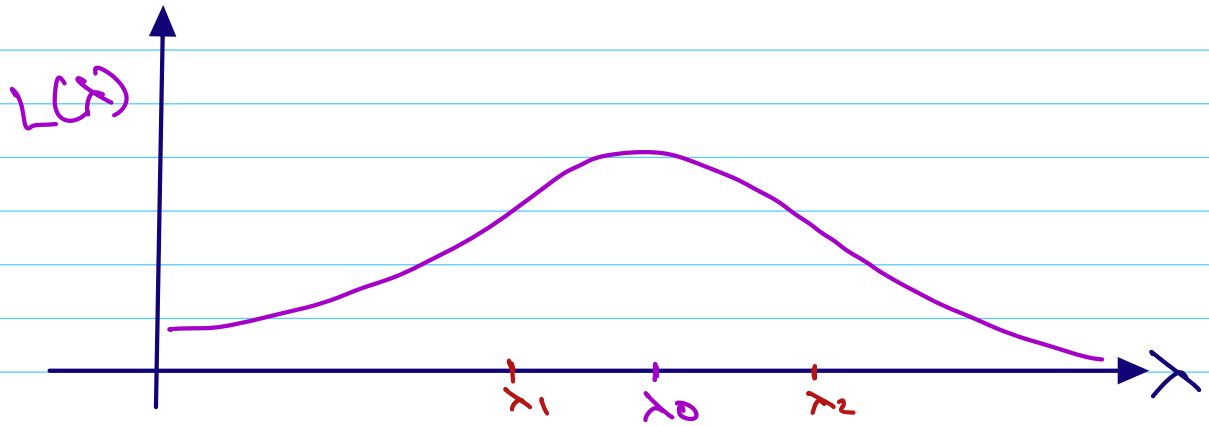


+ The above Likelihood function have very sharp Peak in it. meaning at some value λ_0, a true underlying value of our Parameter λ, we were very likely to get our data.

* Our data $x_1, x_2, \ldots, x_n$ is inconsistant with λ being $\lambda < \lambda_1$ or $\lambda > \lambda_2$. But it extremly likely we would have our data if our $\lambda = \lambda_0$

$\Rightarrow$ The true Parameter λ^* will be very close to λ_0.

Now Imagine another Likelihood function, where the distribution is very Broad



in the above likelihood function, λ_0 is where the Peak Likelihood is i.e

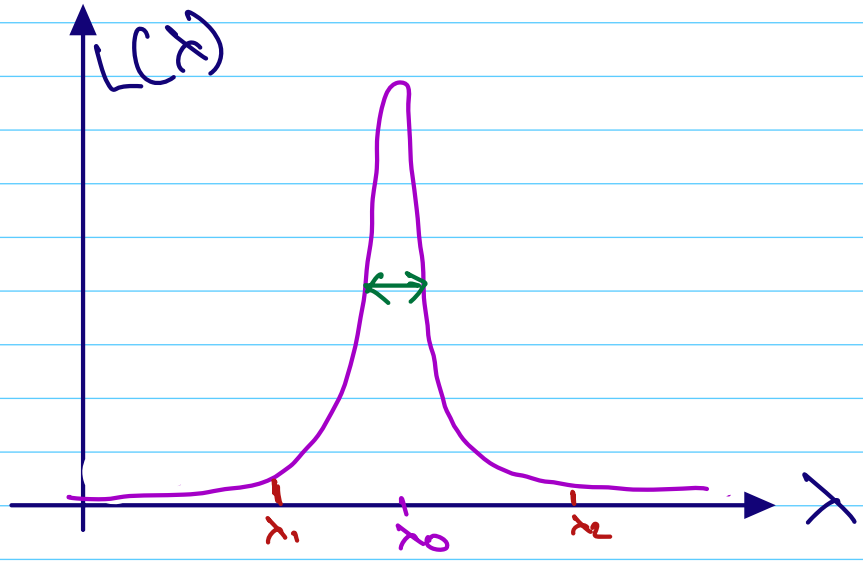
$$\lambda_0 = \underset{\lambda}{\operatorname{argmax}}\ L(\lambda \mid \text{Data})$$

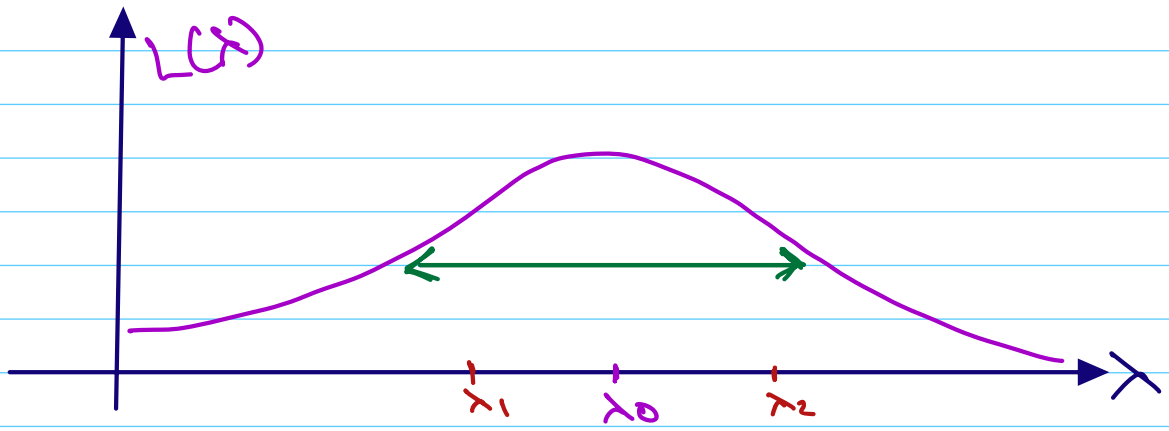
$\Rightarrow$ There is good chance that the data is generated using λ_0, But there is also a good deal of likelihood at some distance away from λ_0, saying that while we are most likely to recover data with the value λ_0, there is a good deal of Likelihood we will acquire data with another values, even if the λ_0 is the true value.

* So, we are looking here some way to quantify the difference b/w the above Likelihood function's. and ask given a set of measurement's, (Data) how constraining

are we on the underlying theory?

- So, the clear sign we are looking for is the width of this likelihood function or some measure of how fast it falls off from the maximum.





if it fall's off very quickly, then our data is relatively constraining, however if it fall's off very slowly, then our data doesn't do much to constrain the underlying theory.

* So, to come up with a more quantitative frame work for it. we will taylor expand the Likelihood function, around its Peak

$$L(\lambda) \approx L(\lambda_0) + \frac{\partial}{\partial\lambda} L(\lambda)\Big|_{\lambda=\lambda_0} (\lambda-\lambda_0)$$

$$+ \frac{1}{2}\frac{\partial^2}{\partial\lambda^2} L(\lambda)\Big|_{\lambda=\lambda_0} (\lambda-\lambda_0)^2$$

we know that $\frac{\partial L}{\partial\lambda}\Big|_{\lambda=\lambda_0} = 0$ because it's Peak.

$$\Rightarrow \quad L(\lambda) \approx L(\lambda_0) + \frac{1}{2} \frac{\partial^2 L}{\partial \lambda^2}\Big|_{\lambda=\lambda_0} (\lambda-\lambda_0)^2$$

we are approximating our likelihood function as parabola. This is generally not a good approximation.

The likelihood can have any shape, so better approximation is to look at 2nd derivative of log-likelihood.

De-tour — will come back after some analysis.

why Parabola approximation fail's, and Gaussian Approximation works.

(1) The likelihood is always positive, but a parabola is not.

if we expand the likelihood function $L(\theta)$ directly around its maximum $\hat{\theta}$:

$$L(\theta) \approx L(\hat{\theta}) + \frac{1}{2} L''(\hat{\theta})(\theta - \hat{\theta})^2$$

(1) Since $L''(\hat{\theta}) < 0$, this is downward opening Parabola.

(2) But this approximation will eventually become negative for sufficiently large $|\theta - \hat{\theta}|$, impossible for Likelihood, because $L(\theta) \geq 0$

So the Parabola approximation is only valid infinitesimally close to the maximum, and has no Probabilistic interpretation.

(2) The Log-Likelihood is the right Object to expand.

The log-Likelihood $\ell(\theta) = \log L(\theta)$

is a smooth, concave function, and taylor-expanding it ensures that the exponentiated form remain positive and normalizable

$$\ell(\theta) = \log L(\theta)$$

Peak $\Leftarrow \hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ L(\theta \mid \text{Data}) = \hat{\theta}_n^{\text{MLE}}$

$\hat{\theta}$ is MLE

$$\ell(\theta) \approx \ell(\hat{\theta}) + \dot{\ell}(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}\ddot{\ell}(\hat{\theta})(\theta - \hat{\theta})^2$$

Exponentiate!

$$\Rightarrow L(\theta) = e^{\ell(\theta)} \approx e^{\ell(\hat{\theta}) + \frac{1}{2}\ddot{\ell}(\hat{\theta})(\theta - \hat{\theta})^2}$$

$$\Rightarrow L(\theta) \approx e^{\ell(\hat{\theta})} \exp\left\{-\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2\right\}$$

where $I(\hat{\theta}) = -\ddot{\ell}(\hat{\theta})$ is the

Observed Fisher information.

This gives a Gaussian kernel, always positive, has a clear Probabilistic interpretation.

Expanding log-likelihood is equivalent to assuming: "Around the peak, the information content changes quadratically"

This means:

- The likelihood drops off exponentially with squared distance from the maximum.
- The exponential decay reflects how evidence weakens as we move

away from the best parameter

* The curvature (Fisher information) determines the rate of decay, which directly measures how strongly data constrain θ.

$$L(\theta) \approx e^{\ell(\hat{\theta})} \cdot e^{\frac{1}{2}(\theta-\hat{\theta})^2 \ell''(\hat{\theta})}$$

Define $$I(\hat{\theta}) = -\ell''(\hat{\theta})$$

(That's the observed Fisher information, the curvature at the peak).

$$\Rightarrow L(\theta) \approx e^{\ell(\hat{\theta})} \cdot e^{-\frac{1}{2} I(\hat{\theta})(\theta-\hat{\theta})^2}$$

Recognize the Gaussian form

The term $e^{-\frac{1}{2} I(\hat{\theta})(\theta-\hat{\theta})^2}$

is exactly the kernel of a Normal distribution with

$$\theta \sim \mathcal{N}\left(\hat{\theta}, \frac{1}{I(\hat{\theta})}\right)$$

So around its maximum, the likelihood behaves like a Gaussian centered the MLE with variance equal to the inverse of the curvature.

Back to our content

in multi parameter space, the

2nd derivative is Hessian

$$\begin{bmatrix} \frac{\partial^2 L}{\partial \lambda^2} & \frac{\partial L}{\partial \lambda}\frac{\partial L}{\partial \beta} \\ \frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \lambda} & \frac{\partial^2 L}{\partial \beta^2} \end{bmatrix} = -F$$

This we define $-F$, where F is the Fisher or Curvature matrix

* The Fisher matrix is sometimes called Curvature matrix, because it is 2nd derivative of log-likelihood function

* It tells how curved the likelihood function is around it's maximum.

So the bigger the values in the

fisher matrix, the more curved it is, meaning the more peaky it is, meaning the more constraing our data is for that particular parameter.

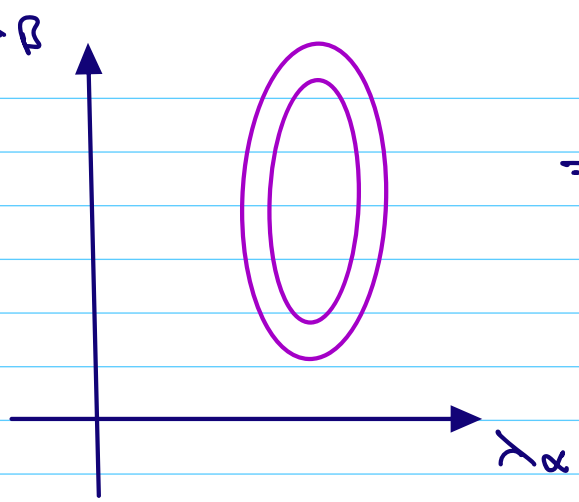
Another expression in statistic's, the Co-variance matrix, which is related to the fisher matrix as follow's.

$$\Rightarrow \quad F_{\alpha\beta}^{-1} = C_{\alpha\beta}$$

Co-variance is inverse of Fisher matrix.

$$C_{\alpha\beta} = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\beta\alpha} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}$$

(i)



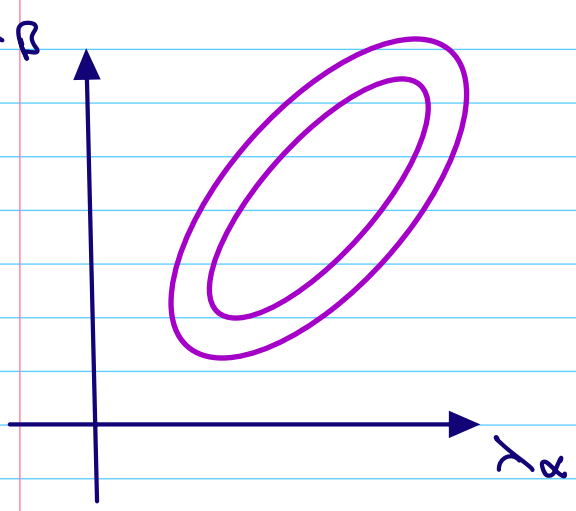
$\Rightarrow$

$$\begin{bmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\beta^2 \end{bmatrix}$$

$\sigma_\beta^2 > \sigma_\alpha^2$

$\sigma_{\alpha\beta} = \sigma_{\beta\alpha} = 0$ (no dependency)

(ii)



the covariance.

another interpretation:

$$x \sim f_x(x|\theta) \quad \theta \in \Theta$$

we don't know the true parameter θ^*

$\Rightarrow$ let's assume we observed one realization $x \sim f_x(x|\theta^*)$ and want to know how much this observation tells us about the true parameter.