ADA 07 - 9am Tue 27 Sep 2022

Maximum Likelihood Estimation
Error bars are Model parameters
Fitting Poisson Data
Noise Model Parameters

Conditional Probabilities
Bayes Theorem
Bayesian Inference

135

Example: Correct the Bias in $(S^2)^{1/2}$

Define $y(x) = x^b$,

Derivatives: $y'(x) = b\,x^{b-1}, \quad y''(x) = b(b-1)\,x^{b-2}$

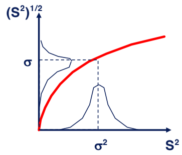
Evaluate the bias:

$$\left\langle \left(S^2\right)^b \right\rangle = y\left(\left\langle S^2 \right\rangle\right) + \frac{y''\left(\left\langle S^2 \right\rangle\right)}{2}\,\mathrm{Var}\left(S^2\right) + \dots$$

$$= y\left(\sigma^2\right) + \frac{y''\left(\sigma^2\right)}{2}\,\frac{2\sigma^4}{N-1} + \dots$$

$$= \sigma^{2b} + \frac{b(b-1)\sigma^{2(b-2)}}{2}\,\frac{2\sigma^4}{N-1} + \dots \quad = \sigma^{2b}\left(1 + \frac{b(b-1)}{N-1} + \dots\right)$$

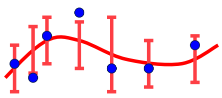
$$\left\langle \left(S^2\right)^{p/2} \right\rangle = \sigma^p \left(1 + \frac{p(p-2)}{4(N-1)} + \dots\right)$$



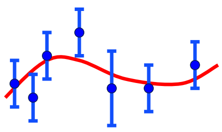
Bias-corrected: $\bar{S} \equiv \dfrac{\sqrt{S^2}}{\left(1 + \dfrac{p(p-2)}{4(N-1)}\right)^{1/p}} \qquad \left\langle \bar{S}^p \right\rangle = \sigma^p$

136

Error Bars live with the Model



Not with the Data



Usually the distinction is unimportant.

But sometimes ***it is important***.

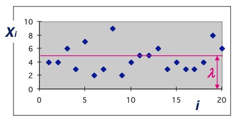
137

Error bars live with the model, not the data!

Example: **Poisson data:**

$$\mathrm{Prob}(x = n \mid \lambda) = \frac{\lambda^n\, e^{-\lambda}}{n!} \qquad n = 0, 1, 2, \dots$$

$$\langle X_i \rangle = \lambda, \qquad \sigma^2(X_i) = \lambda$$



How to attach error bars to the data points?

The **wrong way**: If $\sigma(X_i) = \sqrt{X_i}$, then $1/\sigma^2 = \infty$ when $X_i = 0$

and $\hat{X} \equiv \dfrac{\sum_i X_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} = \dfrac{0 \cdot \infty}{\infty} = 0$, clearly wrong !

Assigning $\sigma(X_i) = \sqrt{X_i}$ gives a **downward bias**. Points lower than average by chance are given smaller error bars, and hence more weight than they deserve.

The **right way**:
Assign $\sigma = \sqrt{\lambda}$, where λ = mean count rate ***predicted by the model.***

138

Maximum Likelihood (ML) Estimation

Likelihood of parameters α for a given dataset:

$$L(\alpha) \equiv P(X \mid \alpha) = P(X_1 \mid \alpha) \times P(X_2 \mid \alpha) \times \dots \times P(X_N \mid \alpha)$$

$$\equiv \prod_{i=1}^{N} P(X_i \mid \alpha)$$

Example: **Gaussian errors:**

$$P(X_i \mid \alpha) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left\{ -\frac{1}{2}\left(\frac{X_i - \mu_i(\alpha)}{\sigma_i}\right)^2 \right\}$$

$$L(\alpha) = \frac{\exp\left\{-\chi^2/2\right\}}{Z_D}, \quad Z_D \equiv (2\pi)^{N/2} \prod_{i=1}^{N} \sigma_i$$

$$\mathrm{BoF} = -2 \ln L = \chi^2 + \sum_i \ln \sigma_i^2 + N \ln(2\pi)$$

To maximise $L(\alpha)$, minimise $\chi^2 + \sum_i \ln \sigma_i^2$

Maximum Likelihood Parameters

α_{ML} satisfies $0 = \dfrac{\partial}{\partial \alpha}[-2 \ln L(\alpha)]$,

$$\mathrm{Var}[\alpha_{\mathrm{ML}}] \approx \frac{2}{\left(\dfrac{\partial^2}{\partial \alpha^2}[-2 \ln L(\alpha)]\right)_{\alpha = \alpha_{ML}}}$$

Generalises χ^2 fitting.

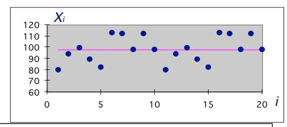
1. For parameters that affect σ
2. For non-Gaussian errors

139

Need ML when Parameters alter Error Bars

- Data points X_i with no error bars:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

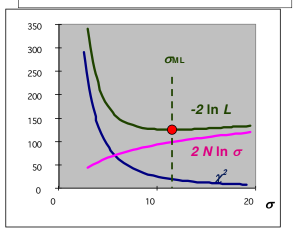


- To find μ, minimise χ^2.
- To find σ, minimising χ^2 fails!

$$\chi^2 \to 0 \quad \text{as } \sigma \to \infty$$

- ML method minimises

$$-2 \ln L = \chi^2 + N \ln \sigma^2$$



140

Need ML to fit low-count Poisson Data

Example : **Poisson data** :

$$P(X = n \mid \lambda) = \frac{e^{-\lambda}\lambda^n}{n!} \qquad n = 0,1,...\infty$$

Likelihood for N Poisson data points :

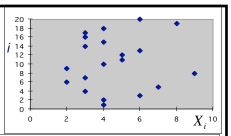
$$L(\lambda) = \prod_{i=1}^{N} P(X_i \mid \lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$$

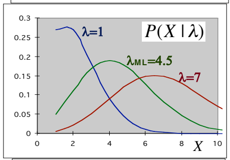
$$\ln L = \sum_i \left(-\lambda + X_i \ln\lambda - \ln X_i! \right)$$

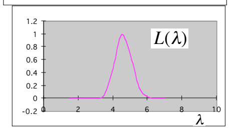
Maximum likelihood estimator of λ :

$$\frac{\partial \ln L}{\partial \lambda} = -N + \frac{1}{\lambda}\sum_i X_i = 0 \quad \text{at} \quad \lambda = \lambda_{ML}$$

$$\therefore \quad \lambda_{ML} = \frac{1}{N}\sum_i X_i .$$







141

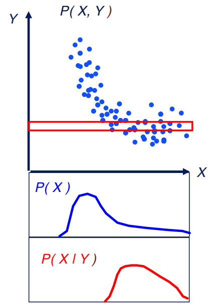
Conditional Probabilities

$P(X,Y)$ = ***joint probability density*** of X and Y
$P(X)$ = projection of $P(X,Y)$ onto X axis.

$$P(X) = \int P(X,Y)\, dY$$

Conditional Probability:
$P(X \mid Y)$ = "probability of X given Y"
= "normalised slice" of $P(X,Y)$
at a **fixed value** of Y.

$$P(X \mid Y) \equiv \frac{P(X,Y)}{P(Y)} = \frac{P(X,Y)}{\int P(X,Y)\, dX}$$

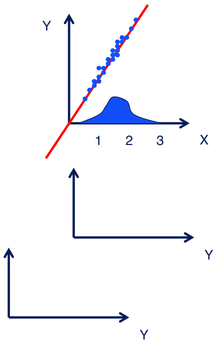


142

Test Understanding

Y = 3 X

X = Gaussian



P(Y I X = 2) = ?

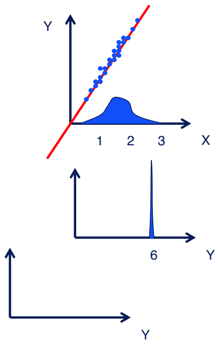
P(Y I X > 2) = ?

143

Test Understanding

Y = 3 X

X = Gaussian



P(Y I X = 2) = ?

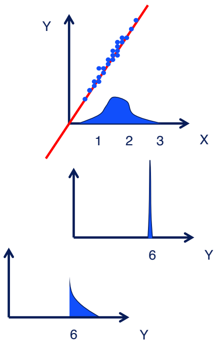
P(Y I X > 2) = ?

144

Test Understanding

Y = 3 X

X = Gaussian



P(Y I X = 2) = ?

P(Y I X > 2) = ?

145

Conditional Probabilities

$P(X)$ = projection onto X axis.
$P(Y)$ = projection onto Y axis.

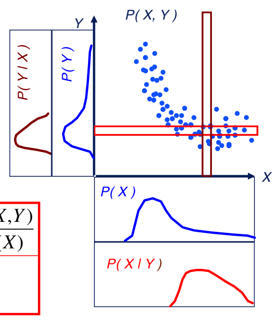
$$P(X) = \int P(X,Y)\, dY$$

$$P(Y) = \int P(X,Y)\, dX$$

Conditional Probability:
$P(X \mid Y)$ = normalised slice at fixed Y
$P(Y \mid X)$ = normalised slice at fixed X

$$P(X \mid Y) \equiv \frac{P(X,Y)}{P(Y)} \qquad P(Y \mid X) \equiv \frac{P(X,Y)}{P(X)}$$

$$P(X,Y) = P(X \mid Y)\, P(Y)$$
$$= P(Y \mid X)\, P(X)$$



146

Bayes' Theorem and Bayesian Inference

Bayes' Theorem: $P(X\,|\,Y) = \dfrac{P(Y\,|\,X)\,P(X)}{P(Y)}$

Since $P(X,Y) = P(X\,|\,Y)\,P(Y) = P(Y\,|\,X)\,P(X)$

then $P(X\,|\,Y) = \dfrac{P(Y\,|\,X)\,P(X)}{P(Y)} = \dfrac{P(Y\,|\,X)\,P(X)}{\int P(Y\,|\,X)\,P(X)\,\mathrm{d}X}$

Bayesian Inference :

$$P(\text{model}\,|\,\text{data}) = \frac{P(\text{data}\,|\,\text{model})\,P(\text{model})}{P(\text{data})}$$

Shows us ***how to change*** our probability distribution P(model) => P(model I data) over various models in light of new data.

147

Inferences depend on Prior, not just Data

Bayesian inference: (M = model, D = data)

Posterior Probability = (Likelihood × Prior Probability) / Evidence

$$P(M\,|\,D) = \frac{P(D\,|\,M)\,P(M)}{P(D)} = \frac{P(D\,|\,M)\,P(M)}{\int P(D\,|\,M)\,P(M)\,\mathrm{d}M}$$

Relative probability of two models M_1 and M_2 :

$$\frac{P(M_1\,|\,D)}{P(M_2\,|\,D)} = \frac{P(D\,|\,M_1)}{P(D\,|\,M_2)} \times \frac{P(M_1)}{P(M_2)} \approx \exp\left(\frac{-\Delta\chi^2}{2}\right) \times \frac{P(M_1)}{P(M_2)}$$

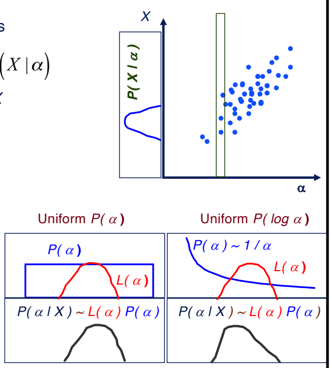
- The ***Likelihood***, P(data I model), is quantified by a ***"badness-of-fit"*** statistic. e.g. P(data I model) ~ exp($-\chi^2/2$)
- The ***Prior***, P(model) expresses your ***prejudice*** (prior knowledge).
- The ***Posterior***, P(model I data), gives your ***inference***, the relative probabilities of different models (parameters), in light of the data.

No absolute inferences ! New data **updates** your prior expectations, but **your conclusions depend also on your prior**.

148

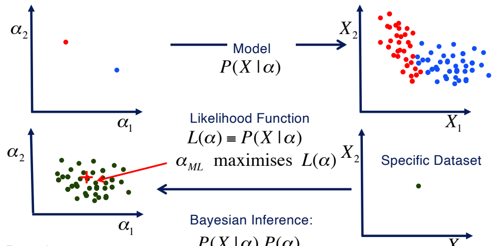
Choice of Prior

- A model for a set of data X depends on model parameters α, and gives the Likelihood $L(\alpha) \equiv P(X\,|\,\alpha)$
- Knowledge of α before measuring X is quantified by the **prior** $P(\alpha)$.
- Choice of prior $P(\alpha)$ is arbitrary, subject to common sense!
- After measuring X, Bayes theorem gives **posterior** : $P(\alpha\,|\,X) \propto P(X\,|\,\alpha)\,P(\alpha)$ $= L(\alpha)\,P(\alpha)$
- Different priors $P(\alpha)$ lead to different **inferences** :



149

Max Likelihood and Bayesian Inference



Bayesian Inference:

Posterior Probability $P(\alpha\,|\,X) = \dfrac{P(X\,|\,\alpha)\,P(\alpha)}{\int P(X\,|\,\alpha)\,P(\alpha)\,\mathrm{d}\alpha} \propto L(\alpha)\,P(\alpha)$

α_{MAP} maximises posterior : $L(\alpha)\,P(\alpha)$ Likelihood updates the Prior.

150

N=1 Gaussian Datum with Uniform Prior

Data : $X \pm \sigma$ Model parameter : μ

Likelihood function :

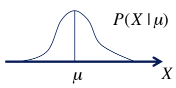
$$L(\mu) \equiv P(X\,|\,\mu) = \frac{e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma}$$

$\mu_{ML} = X$ maximises $L(\mu)$.

Posterior probability :

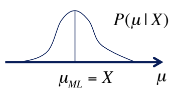
$$P(\mu\,|\,X) = \frac{P(X\,|\,\mu)\,P(\mu)}{P(X)}$$

$$P(X) = \int P(X\,|\,\mu)\,P(\mu)\,\mathrm{d}\mu$$



Uniform prior:

$P(\mu)$ = constant



Maximum Likelihood implicitly assumes a Uniform Prior

151

N=1 Gaussian Datum with Gaussian Prior

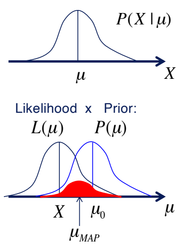
Gaussian Data: $X \pm \sigma$

Likelihood: $L(\mu) \equiv P(X\,|\,\mu) = \dfrac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$

Prior: $P(\mu) = \dfrac{1}{\sqrt{2\pi}\,\sigma_0}\, e^{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2}$

Posterior : $P(\mu\,|\,X) \propto$ Likelihood x Prior

$$L(\mu)\,P(\mu) \propto e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2} \propto \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_{MAP}}{\sigma(\mu_{MAP})}\right)^2\right\}$$



Maximum Posterior (MAP) estimate:

$$\mu_{MAP} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{X}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}, \quad \mathrm{Var}(\mu_{MAP}) = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}.$$

Verify this result.

Same as Optimal Average !

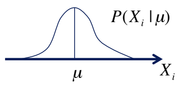
Gaussian prior acts like 1 more data point.

Data "pulls" the probability away from the prior, and vice-versa.

152

N Gaussian Data with Gaussian Prior

Likelihood: $L(\mu) \equiv P(X \mid \mu) = \prod_{i=1}^{N} P(X_i \mid \mu) = \dfrac{\exp\left\{-\frac{1}{2}\chi^2\right\}}{(2\pi)^{N/2} \prod_i \sigma_i}$



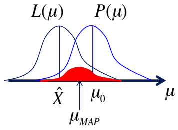
Prior: $P(\mu) = \dfrac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}$

Posterior : $P(\mu \mid X) \propto$ Likelihood x Prior

$$L(\mu)\,P(\mu) \propto \exp\left\{-\frac{\chi^2}{2} - \frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\} \propto \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_{MAP}}{\sigma(\mu_{MAP})}\right)^2\right\}$$

Likelihood x Prior:

$L(\mu)$ $P(\mu)$



Maximum Posterior (MAP) estimate:

$$\mu_{MAP} = \frac{\dfrac{\mu_0}{\sigma_0^2} + \displaystyle\sum_{i=1}^{N} \frac{X_i}{\sigma_i^2}}{\dfrac{1}{\sigma_0^2} + \displaystyle\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}, \quad \sigma^2(\mu_{MAP}) = \frac{1}{\dfrac{1}{\sigma_0^2} + \displaystyle\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}.$$

Same as Optimal Average !

Gaussian prior acts like 1 more data point.

153

Summary:

1. Error bars live with the Model, not with the Data.
2. Bayes Theorem (**Bayesian Inference**)

$$P(\text{Model} \mid \text{Data}) = \frac{P(\text{Data} \mid \text{Model})\,P(\text{Model})}{P(\text{Data})}$$

3. **Maximum Likelihood**, $L(\text{Model}) \equiv P(\text{Data} \mid \text{Model})$

   e.g. for Gaussian Data:

$$BoF = -2\ln L = \chi^2 + \sum_{i=1}^{N} \ln\sigma_i^2 + const$$

4. Minimise χ^2 if Gaussian errors with known σ_i .
5. or Maximise likelihood (e.g. minimise *BoF = -2* ln *L*),
   if error bars unknown, or low-count Poisson data.
6. or full **Bayesian analysis**, including the prior:

   e.g. for Gaussian Data:

$$BoF = -2\ln P(\text{Model} \mid \text{Data}) = \chi^2 + \sum_{i=1}^{N} \ln\sigma_i^2 - 2\ln P(\text{Model}) + const$$

154

Fini -- ADA 07

155