

1. BASICS

Rules of probability theory:

$$\text{prob}(X|I) + \text{prob}(\bar{X}|I) = 1$$

SUM RULE

$$\text{prob}(X, Y|I) = \text{prob}(X|Y, I) \times \text{prob}(Y|I)$$

PRODUCT
RULE

\bar{X} : X is false : proposition

$|$: given (I) ; means I is true

↑ Background information

Bayes' THEOREM:

$$\text{prob}(X|Y, I) = \frac{\text{prob}(Y|X, I) \times \text{prob}(X|I)}{\text{prob}(Y|I)}$$

Marginalization:

$$\text{prob}(X|I) = \int_{-\infty}^{+\infty} \text{prob}(X, Y|I) dY$$

(2)

Bayes' theorem from product rule

$$\text{prob}(Y, X | I) = \text{prob}(Y | X, I) \text{prob}(X | I)$$

$$= \text{prob}(X | Y, I) \text{prob}(Y | I)$$

$$(\text{from } \text{prob}(Y, X | I) = \text{prob}(X, Y | I))$$

consequence for data:

$$\underbrace{\text{prob}(\text{hypothesis} | \text{data}, I)}_{\text{posterior}} \propto \underbrace{\text{prob}(\text{data} | \text{hypothesis}, I)}_{\text{Likelihood}} \times \underbrace{\text{prob}(\text{hypothesis} | I)}_{\text{prior}}$$

prior: state of knowledge (or ignorance)

before analysis of current data

(likelihood: probability of the data given hypothesis
(measurement))

posterior probability: truth of hypothesis in light of new data

$\text{prob}(\text{data} | I)$ omitted: not relevant for parameter estimation \Rightarrow normalization constant

In model selection: EVIDENCE

Marginalization :

first $\text{prob}(X|I) = \text{prob}(X, Y|I) + \text{prob}(X, \bar{Y}|I)$ (*)

proof: $\text{prob}(X, Y|I) = \text{prob}(Y, X|I) = \text{prob}(Y|X, I) \times \text{prob}(X|I)$

hence: $\text{prob}(X, Y|I) + \text{prob}(X, \bar{Y}|I)$

$$= [\text{prob}(Y|X, I) + \text{prob}(\bar{Y}|X, I)] \times \text{prob}(X|I)$$

□

(*): probability of X is true is sum of probability of X AND Y is true plus probability of X AND Y is not true.

Assume set of alternative hypothesis:

$$Y_1, Y_2, \dots, Y_M = \{Y_k\}$$

then $\text{prob}(X|I) = \sum_{k=1}^M \text{prob}(X, Y_k|I)$

(can be derived in similar fashion as previous)

use: $\text{prob}(X, Y_k|I) = \text{prob}(Y_k|X, I) \times \text{prob}(X|I)$

need: $\sum_{k=1}^M \text{prob}(Y_k|X, I) = 1$

NORMALIZATION: need Y_k or mutually exclusive and exhaustive!
(if one Y_k is true all others are false)

Example: in which range is Hubble constant H_0
(continuous intervals, big range) \Rightarrow continuum limit
 $M \rightarrow \infty$

in that case prob \rightarrow probability **DENSITY** function

$$\text{pdf}(X, Y=y | I) = \lim_{\delta y \rightarrow 0} \frac{\text{prob}(X, y \leq Y \leq y + \delta y | I)}{\delta y}$$

\Rightarrow probability of Y in range y_1 and y_2 :

$$\text{prob}(X, y_1 \leq Y < y_2 | I) = \int_{y_1}^{y_2} \text{pdf}(X, Y | I) dY$$

(in future: use prob = pdf)

$$\Rightarrow \int_{-\infty}^{+\infty} \text{prob}(Y | X, I) dY = 1$$

marginalization: important for "NUISANCE" parameters
 \downarrow
of no intrinsic interest

probability: degree -of- belief !

2. PARAMETER ESTIMATION I

(5)

"Mass of Saturn
charge of electron, ..."

2.1. EXAMPLE I: Is this a fair coin?

"observed 4 heads in 11 flips"

Do you think coin is fair?

fair: even 50:50 bet on outcome of flip being
head or tail

if we decide coin was fair, how sure are we?

propositions: $H=0$ and $H=1$ tails or heads
on every flip!

$H=1/2$: fair coin

possible proposition (continuous):

$$0 \leq H < 0.01$$

$$0.01 \leq H < 0.02$$

$$0.02 \leq H < 0.03$$

\vdots

State of knowledge about fairness is how much we believe
these propos. to be true
bias - weighting of coin?

fairness of coin: $\text{prob}(H | \{\text{data}\}, I)$

or better: $\text{prob}(H | \{\text{data}\}, I) dH$

Use Bayes':

$$\text{prob}(H | \{\text{data}\}, I) \propto \text{prob}(\{\text{data}\} | H, I) \times \text{prob}(H | I)$$

(normalization omitted)

prior: $\text{prob}(H | I)$

"very open minded prior":

$$\text{prob}(H | I) = \begin{cases} 1 & 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{UNIFORM PRIOR } (\Delta)$$

assume flips of coin are independent events (I)

probability of obtaining 'R heads in N tosses'

$$\text{prob}(\{\text{data}\} | H, I) \propto H^R (1-H)^{N-R}$$

(*)

(will be derived later) BINOMIAL

product of (Δ) and (*) to ^{get} posterior pdf

PLOT
Lecture 1. R

just prior

N =	0	16	1024
	1	32	2048
	2	64	
	3	128	4096
	4	256	
	8	512	

no evidence for head (or tail)

bias: weighting

2.1.1 Different Priors

a) Gaussian prior around $\bar{H}=0.5$; $\sigma=0.07$
(most coins are fair) - dashed line

$$p(H|I) = \frac{1}{\sqrt{\pi} \sigma} \exp\left[-\frac{1}{2} \frac{(H-\bar{H})^2}{\sigma^2}\right]$$

NORM
NOT NEEDED

b) heavily peaked at "0" and "1"

$$p(H|I) = \begin{cases} \frac{1}{\sqrt{2\pi} \sigma} \left\{ \exp\left[-\frac{1}{2} \frac{H^2}{\sigma^2}\right] + \exp\left[-\frac{1}{2} \frac{(H-1)^2}{\sigma^2}\right] \right\} & 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

choose $\sigma = 0.02$

→ indicates a strange coin!

SHOW LECTURE 1.2.R

← scale to same height! at max!

WE NOTE:

- A FEW SAMPLES DO NOT HELP THE DECISION
- IT TAKES A LOT OF FLIPS TO NOTICE BIAS WEIGHTING
($N=1000$ TO PUT IT INTO $0.2 - 0.3$)
- SOLID AND DOTTED CONVERGES MUCH QUICKER

BOTH SOLID AND DOTTED ARE FAIRLY "FLAT" PRIORS!
⇒ LARGE DEGREE OF IGNORANCE

DASHED: TAKES MORE TO OVERCOME FAIRNESS ASSUMPTION!

2.1.2 SEQUENTIAL OR ONE-STEP ANALYSIS?

8

SET OF DATA $\{D_k\}$

(e.g. N flips of a coin ($k=1,2,\dots,N$))

Bayes' theorem:

$$\text{prob}(H | \{D_k\}, I) \propto \text{prob}(\{D_k\} | H, I) \times \text{prob}(H | I)$$

ONE-STEP PROCESS: CONSIDER DATA COLLECTIVELY

SEQUENTIAL APPROACH:

D_1 : $\text{prob}(H | D_1, I)$ use as prior for analysis of D_2 !

\Rightarrow 2nd posterior; prior for third, ...

EXAMPLE: 2 DATA

$$\Rightarrow \text{prob}(H | D_2, D_1, I) \propto \text{prob}(D_2, D_1 | H, I) \times \text{prob}(H | I)$$

but from Bayes' also:

$$\text{prob}(H | D_2, D_1, I) \propto \text{prob}(D_2 | H, D_1, I) \times \text{prob}(H | D_1, I)$$

(I: assumed data are independent)

\hookrightarrow given H , the value of one flip does not depend on another!

$$\Rightarrow \text{prob}(D_2 | H, D_1, I) = \text{prob}(D_2 | H, I)$$

$$\Rightarrow \text{prob}(H | D_2, D_1, I) \propto \text{prob}(D_2 | H, I) \times \text{prob}(H | D_1, I)$$

□

can be extended: $\text{prob}(H | D_3, D_2, D_1, I) \propto$
 $\text{prob}(D_3 | H, I) \times \text{prob}(H | D_1, D_2, I)$

⇒ BOTH ONE-STEP AND SEQUENTIAL
ANALYSIS GIVE THE SAME ANSWER !

DO NOT USE :

POSTERIOR AS PRIOR FOR SAME
ANALYSIS !



2.2. Reliabilities: best estimates, error-bars and confidence intervals

- ^{posterior} pdf encodes inference about the value of a parameter given data ~~and~~.

↳ want to know best estimate and reliability

best estimate is maximum of posterior pdf:

$$P = \text{prob}(x | \{\text{data}\}, I)$$

best estimate x_0 : $\left. \frac{dP}{dx} \right|_{x_0} = 0$

(also check: $\left. \frac{d^2P}{dx^2} \right|_{x_0} < 0$)

reliability measure: spread of posterior around x_0
to study function in neighborhood of point:

Taylor series

it is better to study logarithm (less "peaky" than pdf)

$$L = \ln [\text{prob}(x | \{\text{data}\}, I)]$$

varies slower with x than pdf

$$L \approx L(x_0) + \frac{1}{2} \left. \frac{d^2L}{dx^2} \right|_{x_0} (x - x_0)^2 + \dots$$

(since $\left. \frac{dL}{dx} \right|_{x_0} = 0$)

$$\hat{=} \left. \frac{dP}{dx} \right|_{x_0} = 0$$

$L(x_0)$: constant \rightarrow not relevant for shape of pdf

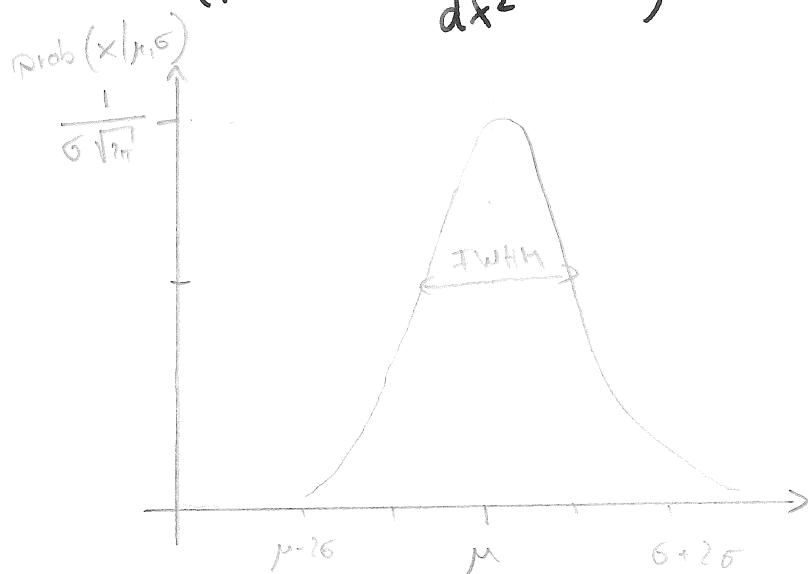
(2)

\Rightarrow Ignoring higher orders

$$\Rightarrow \text{prob}(X|\{\text{data}\}, I) = \underset{\substack{\uparrow \\ \text{NORMALIZATION} \\ \text{CONSTANT}}}{A} \exp\left[\frac{1}{2} \frac{d^2 L}{dx^2} \Big|_{x_0} (x-x_0)^2\right]$$

\Rightarrow approximation by Gaussian

(NOTE: $\frac{d^2 L}{dx^2} < 0$)



$$\text{prob}(x|\mu, \sigma) =$$

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Symmetric about
maximum
width $\sim \sigma$

Gaussian, maximum at $x=\mu$; full width half maximum $\approx 2.35\sigma$

$$\Rightarrow \sigma = \left[-\frac{d^2 L}{dx^2} \Big|_{x_0} \right]^{-1/2}$$

$x=x_0 \pm \sigma$ is where most of the information is

σ : measure of reliability: usually ERROR-BAR

$$\text{prob}(X_0 - \sigma \leq X \leq X_0 + \sigma \mid \{\text{data}\}, I)$$

$$= \int_{X_0 - \sigma}^{X_0 + \sigma} \text{prob}(X \mid \{\text{data}\}, I) dX \approx 0.67$$

probability that true value of X lies within $\pm 1\sigma$ of $X = X_0$ is 67%

Similar: probability $\pm 2\sigma$ of X_0 is 95%

2.2.1 The coin example

(assume uniform prior)

$$\text{prob}(H \mid \{\text{data}\}, I) \propto H^R (1-H)^{N-R}$$

where $0 \leq H \leq 1$

$$\Rightarrow L = \text{constant} + R \ln H + (N-R) \ln (1-H)$$

$$\frac{dL}{dH} = \frac{R}{H} - \frac{N-R}{1-H} \quad \text{and}$$

$$\frac{d^2L}{dH^2} = -\frac{R}{H^2} - \frac{N-R}{(1-H)^2}$$

$$\frac{dL}{dH} \Big|_{H_0} = 0 \Rightarrow \frac{R}{H_0} = \frac{N-R}{1-H_0} \Rightarrow H_0 = \frac{R}{N}$$

$$\Rightarrow \frac{d^2L}{dH^2} \Big|_{H_0} = -\frac{N}{H_0(1-H_0)}$$

$$\Rightarrow \sigma = \sqrt{\frac{H_0(1-H_0)}{N}}$$

(4.)

SINCE H_0 does not vary a lot after
a moderate amount of data $\Rightarrow H_0 = \text{const}$

$$\Rightarrow \sigma \sim \frac{1}{\sqrt{N}}$$

Lecture 2.2

write σ and H_0

2.2.2 Asymmetric posterior pdfs

for example low no. of tosses of coin
maximum still gives best estimate
concept of error bar not appropriate

confidence interval

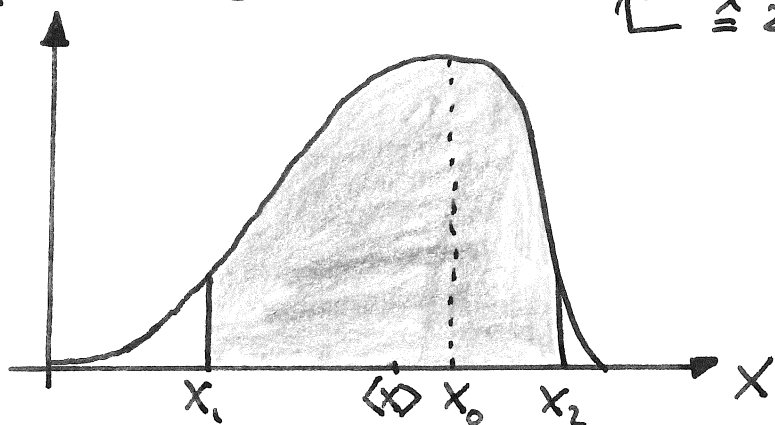
shortest interval that encloses 95% of the area

Find X_1, X_2

$$\text{prob}(X_1 \leq X < X_2 | \{\text{data}\}, I) = \int_{X_1}^{X_2} \text{prob}(X | \{\text{data}\}, I) dX \approx 0.95$$

with $X_2 - X_1$ as small as possible

$X_1 \leq X < X_2$ shortest 95% confidence interval
 $\uparrow \approx 2\sigma$ reasonable



best estimate: maximum: most probable value
 mean or expectation value takes into account
 skewness of pdf

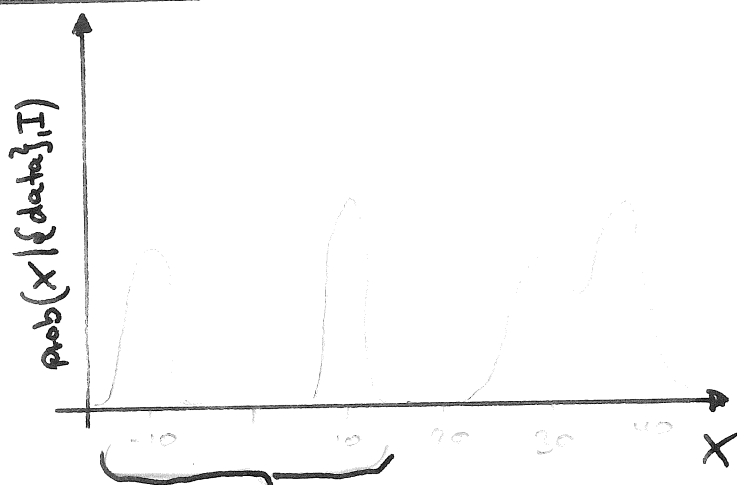
$$\langle X \rangle = \int X \text{prob}(X|\{\text{data}\}, I) dX$$

for normalized
 distribution

$$(E(X); \bar{x})$$

for symmetric distributions: $\langle X \rangle = X_0$

2.2.3 MULTIMODAL POSTERIOR



• What is best estimate?



best estimate and
 error-bar are only
 means to summarize
 pdf → This is some-
 times not possible

bimodal mean $\langle X \rangle = 0$

$X = -10 \pm 3$
 $X = 10 \pm 2$

for bimodal
 two numbers
 + error bar

2.3 Example 2: Gaussian noise and averages

(6)

- often Gaussian is used to theoretically model noise (see later: central limit theorem)

$$\text{prob}(x_k | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_k - \mu)^2}{2\sigma^2}\right]$$

μ : true value of parameter

σ : measure of error

given $\{x_k\}$ what is best estimate of μ and confidence?

assume (for now) : σ is known

$$\text{prob}(\mu | \{x_k\}, \sigma, I) \propto \text{prob}(\{x_k\} | \mu, \sigma, I) \times \text{prob}(\mu | \sigma, I)$$

if x_k are independent

$$\text{prob}(\{x_k\} | \mu, \sigma, I) = \prod_{k=1}^N \text{prob}(x_k | \mu, \sigma, I)$$

uniform pdf for prior:

$$\text{prob}(\mu | \sigma, I) = \text{prob}(\mu | I) = \begin{cases} A & \mu_{\min} \leq \mu \leq \mu_{\max} \\ 0 & \text{otherwise} \end{cases}$$

with $A = \frac{1}{\mu_{\max} - \mu_{\min}}$

$$\Rightarrow L = \ln [\text{prob}(\mu | \{x_k\}, \sigma, I)]$$

$$= \text{constant} - \sum_{k=1}^N \frac{(x_k - \mu)^2}{2\sigma^2}$$

↑

terms not involving μ

Zero outside μ_{\min} and μ_{\max} !

best estimate: $\frac{dL}{d\mu} \Big|_{\mu_0} = 0$

$$\Leftrightarrow \sum_{k=1}^N \frac{x_k - \mu_0}{\sigma^2} = 0$$

$$\Leftrightarrow \sum_{k=1}^N x_k = \sum_{k=1}^N \mu_0 = N\mu_0$$

$$\Rightarrow \mu_0 = \frac{1}{N} \sum_{k=1}^N x_k \quad (\text{arithmetic average})$$

confidence: $\frac{d^2L}{d\mu^2} \Big|_{\mu_0} = - \sum_{k=1}^N \frac{1}{\sigma^2} = - \frac{N}{\sigma^2}$

$$\Rightarrow \mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}$$

$$(\text{again } \sim \frac{1}{\sqrt{N}})$$

if error bar allows values outside $\mu_{\min} - \mu_{\max}$
then use posterior pdf with cut-offs!

in that case: prior as important as data !

2.3.1 Data with different size error bars

(8)

$$\text{prob}(x_k | \mu, \sigma_k) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp \left[-\frac{(x_k - \mu)^2}{2\sigma_k^2} \right]$$

$$\Rightarrow L = \ln [\text{prob}(\mu | \{x_k\}, \{\sigma_k\}, I) = \text{const.} - \sum_{k=1}^N \frac{(x_k - \mu)^2}{2\sigma_k^2}]$$

$$\Rightarrow \mu_0 = \frac{\sum_{k=1}^N w_k x_k}{\sum_{k=1}^N w_k}$$

$$\text{with } w_k = \frac{1}{\sigma_k^2}$$

weighted average

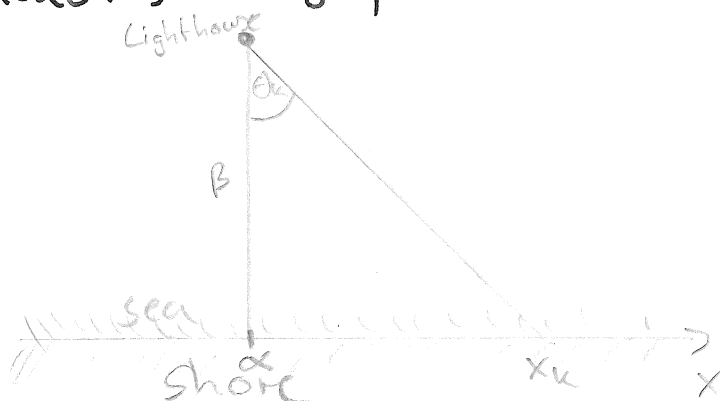
$$\mu = \mu_0 \pm \left(\sum_{k=1}^N w_k \right)^{-1/2}$$

(2nd derivative)

2.4. EXAMPLE: Lighthouse problem

"A lighthouse is in the sea, off a piece straight coast line at position α along shore and distance β out at sea. It emits short (collimated) flashes at random intervals \Rightarrow random

azimuths (angle)



The position x_k of the flashes is recorded

Where is the lighthouse?

uniform prior:

$$\text{prob}(\theta_k | \alpha, \beta, I) = \frac{1}{\pi}$$

$$-\frac{\pi}{2} \leq \theta_k \leq \frac{\pi}{2}$$

with $\tan \theta_k = \frac{x_k - \alpha}{\beta}$

will see later (changing variables)

$$\text{prob}(x_k | \alpha, \beta, I) = \frac{\beta}{\pi [\beta^2 + (x_k - \alpha)^2]}$$

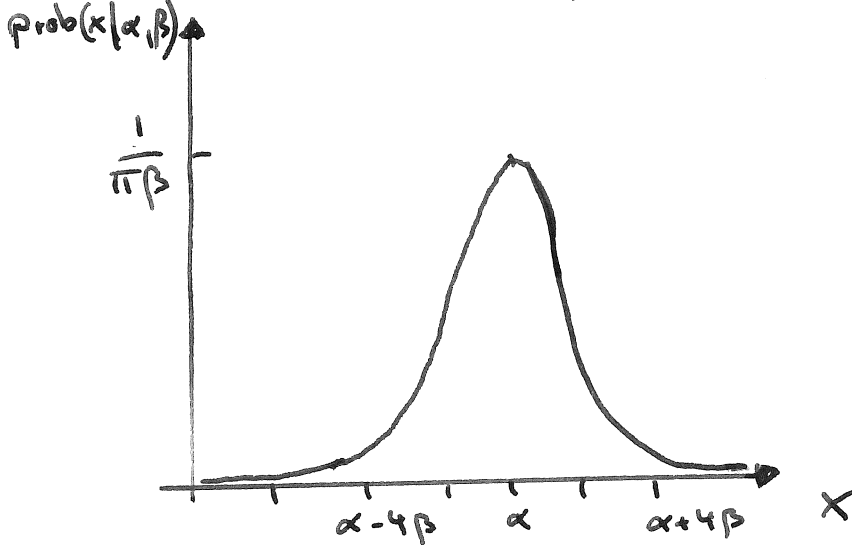
(prob that k-th flash is measured at x_k)

Cauchy - distribution

in physics Lorentzian!

Symmetric about maximum $x_k = \alpha$

and FWHM: 2β



Estimate α and $\beta \rightarrow$ two-parameter problem \rightarrow next week
(for the moment assume β is known!)

$$\Rightarrow \text{prob}(\alpha | \{x_k\}, \beta, I)$$

$$\stackrel{\text{Bayes!}}{=} \text{prob}(\{x_k\} | \alpha, \beta, I) \times \text{prob}(\alpha | \beta, I)$$

unif form prior for α

$$\text{prob}(\alpha | \beta, I) = \text{prob}(\alpha | I) = \begin{cases} A & \alpha_{\min} \leq \alpha \leq \alpha_{\max} \\ 0 & \text{otherwise} \end{cases}$$

$$A = \frac{1}{\alpha_{\max} - \alpha_{\min}}$$

$$\text{prob}(\{x_k\} | \alpha, \beta, I) = \prod_{k=1}^N \text{prob}(x_k | \alpha, \beta, I)$$

$$\Rightarrow L = \ln [\text{prob}(\alpha | \{x_k\}, \beta, I)]$$

$$= \text{const.} - \sum_{k=1}^N \ln [\beta^2 + (x_k - \alpha)^2]$$

all terms not involving α

(no worry about cut-off. Assumption)

$$\frac{dL}{d\alpha} \Big|_{\alpha_0} = 2 \sum_{k=1}^N \frac{x_k - \alpha_0}{\beta^2 + (x_k - \alpha_0)^2} = 0$$

hard to solve analytically

Look for maximum ^{of pdf} "numerically"

numerically easier to work with L
find L_{\max} first and normalize peak of posterior pdf to "1".

Lecture 2b.R

posterior broad
for few data
points

2.4.1 CENTRAL LIMIT THEOREM

CLT

(11)

For Gaussian best estimate is arithmetic average

However for Cauchy average value is NOT good estimate!

Usually error goes down $1/\sqrt{N}$

Cauchy distribution $\rightarrow \sigma$ infinite
 μ not defined

CLT breaks down here!

INTERLUDE : STATISTICAL DATA ANALYSIS
GLEN COWAN
SECTION 2

EXAMPLE OF PROBABILITY FUNCTIONS

BINOMIAL AND ~~MULTIMODAL~~ ~~MULTIMODAL~~ DISTRIBUTIONS

N trials, each two possibilities (success and failure)
probability of success constant p

discrete random variable $r \hat{=}$ number of successes
if one repeats experiment many times with N trials
we would get r distributed binomial

probability of success: p
 probability of failure: $1-p$

trials independent!

example 5 trials: success, success, failure, success, failure

$$\Rightarrow p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3 (1-p)^2$$

$$\Rightarrow \text{in general: } p^r (1-p)^{N-r}$$

(order not important)

number of sequences having r successes in N trials

$$\frac{N!}{r! (N-r)!}$$

$$\Rightarrow \text{prob: } f(r; N, p) = \frac{N!}{r! (N-r)!} p^r (1-p)^{N-r}$$

$$r = 0, \dots, N$$

expectation value of r

$$E[r] = \sum_{r=0}^N r \frac{N!}{r! (N-r)!} p^r (1-p)^{N-r} = Np$$

variance $V[r] = E[r^2] - (E[r])^2 = Np(1-p)$

Lecture 2. R

POISSON DISTRIBUTION

Binomial distribution : limit large N
 p small

$$\nu = N \cdot p \text{ constant}$$

$$\Rightarrow f(r; \nu) = \frac{\nu^r}{r!} e^{-\nu}$$

Poisson distribution for random variable r
 with $r = 0, 1, 2, \dots, \infty$

Lecture 2.2

$$\frac{\nu^r}{r!} = \nu \frac{\nu^{r-1}}{(r-1)!} = \nu \sum_{r=1}^{\infty} \frac{\nu^{r-1}}{(r-1)!} = \nu e^{\nu}$$

Expectation value : $E[r] = \sum_{r=0}^{\infty} r \frac{\nu^r}{r!} e^{-\nu} = \nu$

$$V[r] = \sum_{r=0}^{\infty} (r - \nu)^2 \frac{\nu^r}{r!} e^{-\nu} = \nu$$

(for large $\nu \rightarrow$ Gaussian!)

EXPONENTIAL DISTRIBUTION

continuous variable $0 \leq x < \infty$

$$f(x; \lambda) = \frac{1}{\lambda} e^{-x/\lambda}$$

$$E[x] = \frac{1}{\lambda} \int_0^{\infty} x e^{-x/\lambda} dx = \lambda$$

$$V[x] = \frac{1}{\lambda} \int_0^{\infty} (x - \lambda)^2 e^{-x/\lambda} dx = \lambda^2$$

Lecture 2e.R

(example: decay time of unstable particle)

Log-normal DISTRIBUTION



CONTINUOUS VARIABLE y GAUSSIAN
with mean μ and variance σ^2

$\Rightarrow x = e^y$ follows a log-normal distribution

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} \frac{1}{x} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

$$E[x] = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$$

$$V[x] = \exp(2\mu + \sigma^2) [\exp(\sigma^2) - 1]$$

(NOTE: μ and σ^2 are NOT mean and variance!)

Lecture 2 f.R

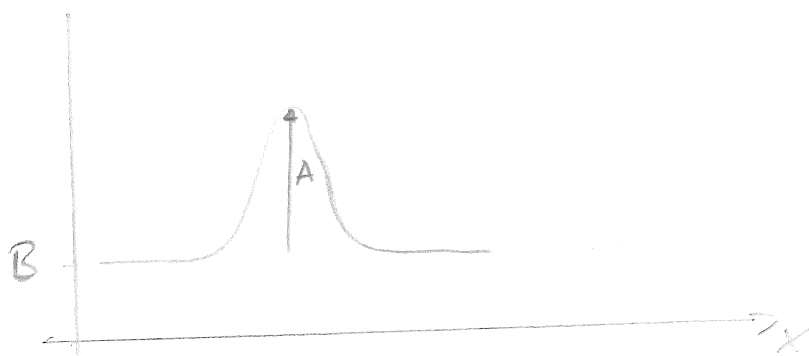
3 Parameter Estimation II

• generalization of previous lecture to several parameters

- generalization of error bars: include correlations
- marginalization: "unwanted" variables

3.1. EXAMPLE 4: Amplitude of a signal in the presence of background

e.g. emission spectrum of galaxy contaminated by stray light from night sky



here: background flat : unknown magnitude B

Signal: amplitude A (known shape and position)

Data: integer valued; e.g. counts of photons

$\{N_k\}$ measured at experimental setting $\{x_k\}$ (for example wavelength)

take peak to be Gaussian at x_0 ; width w

(2)

$$D_k = n_0 \left[A e^{-\frac{(x_k - x_0)^2}{2w^2}} + B \right]$$

n_0 : amount of time for which measurement were made

Assume the number ^{count} N proportional to D_k

Poisson distribution:

$$\text{prob}(N|D) = \frac{D^N e^{-D}}{N!}$$

Lecture 3a

$$n_0 = 1.7 ; 0, \dots, 8$$

$$n_0 = 17.5 ; 0, \dots, 20$$

$$\langle N \rangle = \sum_{N=0}^{\infty} N \text{prob}(N|D) = D$$

for each datum D_k

$$\text{prob}(N_k | A, B, I) = \frac{D_k^{N_k} e^{-D_k}}{N_k!}$$

I : includes relation between D_k and A and B

x_0, w and n_0 are given

for independent data

$$\text{prob}(\{N_k\} | A, B, I) = \prod_{k=1}^M \text{prob}(N_k | A, B, I)$$

(3)

$$\text{prob}(A, B | \{N_k\}, I) \propto \text{prob}(\{N_k\} | A, B, I) \times \text{prob}(A, B | I)$$

$$\text{prob}(A, B | I) = \begin{cases} \text{constant} & \text{for } A \geq 0 \text{ and } B \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(assume A_{\max} and B_{\max} are sufficiently large to not impose a cut-off on the posterior)

$$\Rightarrow L = \ln[\text{prob}(A, B | \{N_k\}, I)]$$

$$= \text{constant} + \sum_{k=1}^M [N_k \ln(D_k) - D_k]$$

↑
all terms which do not include A, B

2.3.5 w. 0.1 - 1.0

Minimize L or maximize prob
reliability: width or sharpness of peak

Lecture 3b.R

1. $x_{\min} = -7$ $n_0 = 33.3$
2. $x_{\min} = -7$ $n_0 = 3.3$
(prior provides cut-off)
3. $x_{\min} = 14$ $n_0 = 33.3$
only B (improved w/ A
(range 20-25, 20 background))
4. $x_{\min} = 3.5$ $n_0 = 33.3$
range restricted
⇒ strong correlation

3.1.1 MARGINAL DISTRIBUTIONS

- 2-dim DISTRIBUTION DESCRIBES COMPLETELY JOINT INFERENCE

- OFTEN NOT INTERESTED in background
 \uparrow
 NOISANCE PARAMETER

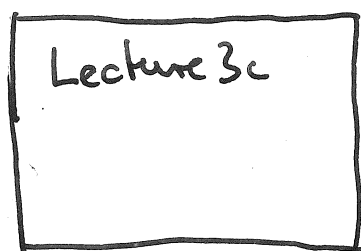
want to know probability of amplitude, irrespective of background:

$$\text{prob}(A | \{N_k\}, I) = \int_0^{\infty} \text{prob}(A, B | \{N_k\}, I) dB$$

MARGINALIZATION

- if interested in background:

$$\text{prob}(B | \{N_k\}, I) = \int_0^{\infty} \text{prob}(A, B | \{N_k\}, I) dA$$



Same as previous
conditional
 $B=2$

- NOTE DIFFERENCE BETWEEN MARGINAL AND CONDITIONAL prob

$$\text{prob}(A | \{N_k\}, B, I)$$

MARGINAL: ignorance of ~~prob~~ value of B

CONDITIONAL: magnitude of B determined

- o For very wide range experiment, little gain from separate calibration of B , for narrow range, the opposite

(5)

generalize if position (x_0) and shape (w) of peak are not well known:

$$\text{prob}(A, B | \{N_k\}, I) = \iint (A, B, w, x_0 | \{N_k\}, I) dw dx_0$$

with Bayes' theorem:

$$\text{prob}(A, B, w, x_0 | \{N_k\}, I) \propto \text{prob}(\{N_k\} | A, B, w, x_0, I) \times \text{prob}(A, B, w, x_0 | I)$$

$$\text{prob}(A, B, w, x_0 | I) = \text{prob}(A, B | I) \times \text{prob}(x_0, w | I)$$

if we would know already shape and position

$$\text{prob}(x_0, w | I) = \delta(w - 2.12) \delta(x_0)$$

fixed w $x_0 = 0$

\Rightarrow integral easy to evaluate

\hookrightarrow in general numerical or analytical integration required

3.1.2 BINNING OF DATA

6

previous histograms:

number counts detected in channels of finite width

one should actually write:

$$D_k = \int_{x_k - \Delta/2}^{x_k + \Delta/2} n_0 \left[A \exp\left(-\frac{1}{2} \frac{(x-x_0)^2}{w^2}\right) + B \right] dx$$

(assuming all measurements have same width)

for small bin-width Δ :

$$D_k \approx n_0 \left[A \exp\left(-\frac{1}{2} \frac{(x-x_0)^2}{w^2}\right) + B \right] \Delta$$

(hence Δ can be absorbed in n_0)

↳ n_0 represents amount of time AND collecting area

is there any thing gained if bins are narrower

Lecture 3d

by = 0.25!

4 times narrower

also $\frac{n_0}{4}$

use routines b, c

→ almost identical

BUT: TOO COARSE BINNING CAN DESTROY USEFUL INFORMATION!

(one may be unable to distinguish signal from background)

later: optimal binning strategies!

3.2. Reliabilities: best estimates, correlations and error-bars

7

o optimal estimate: maximum of posterior pdf
"parameters" of interest: $\{X_j\}$

$$P = \text{prob}(\{X_j\} | \{\text{data}\}, I)$$

best estimate: $\{X_{0j}\}$

$$\left. \frac{\partial P}{\partial X_i} \right|_{\{X_{0j}\}} = 0$$

with $i = 1, 2, \dots$

(also need to ensure maximum)

$$L = \ln[\text{prob}(\{X_j\} | \{\text{data}\}, I)]$$

NOW FOR TWO PARAMETERS (X_0, Y)

$$\left. \frac{\partial L}{\partial X} \right|_{X_0, Y_0} = 0$$

$$\left. \frac{\partial L}{\partial Y} \right|_{X_0, Y_0} = 0$$

$$L = \ln[\text{prob}(X, Y | \{\text{data}\}, I)]$$

RELIABILITY: spread of posterior pdf

TAYLOR EXPANSION:

$$L = L(X_0, Y_0) + \frac{1}{2} \left[\frac{\partial^2 L}{\partial X^2} \Big|_{X_0, Y_0} (X - X_0)^2 + \frac{\partial^2 L}{\partial Y^2} (Y - Y_0)^2 + 2 \frac{\partial^2 L}{\partial X \partial Y} \Big|_{X_0, Y_0} (X - X_0)(Y - Y_0) \right] + \dots$$

(2-dim version of previous)

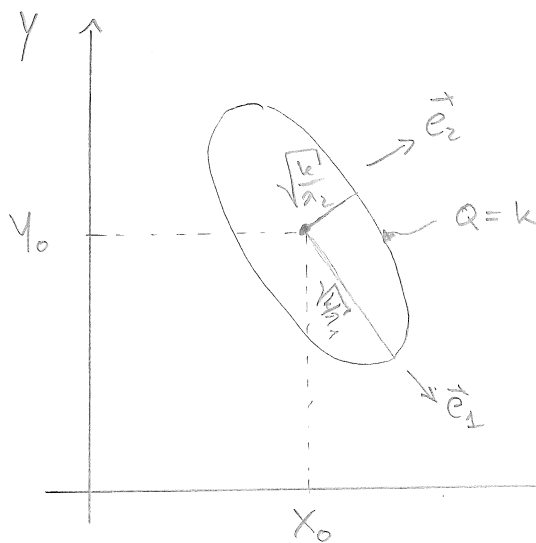
MATRIX NOTATION

$$Q = (X - X_0 \ Y - Y_0) \begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} X - X_0 \\ Y - Y_0 \end{pmatrix}$$

$$A = \frac{\partial^2 L}{\partial X^2} \Big|_{X_0, Y_0}$$

$$B = \frac{\partial^2 L}{\partial Y^2} \Big|_{X_0, Y_0}$$

$$C = \frac{\partial^2 L}{\partial X \partial Y} \Big|_{X_0, Y_0}$$



CONTOUR OF
CONSTANT $Q = k$
(posterior pdf constant)

ELLIPSE CENTRED
AT (X_0, Y_0)

ORIENTATION, ECCENTRICITY : GIVEN BY A, B, C

principal axes : eigenvectors of 2-nd derivative matrix:

$$\begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

TWO EIGENVALUES λ_1 and λ_2 inversely related to
square of width along principal axes

FOR MAXIMUM : BOTH λ_1 and λ_2 must be negative:

$$\Rightarrow A < 0, B < 0, AB > C^2$$

IF ELLIPSE ALIGNED WITH AXES EASY TO OBTAIN RELIABILITY IN x AND y .

ASSUME y IS NUISANCE PARAMETER

$$\text{prob}(X | \{\text{data}\}, I) = \int_{-\infty}^{+\infty} \text{prob}(X, y | \{\text{data}\}, I) dy$$

UNDER TAYLOR EXPANSION (GAUSSIAN APPROX.)

$$\text{prob}(X | \{\text{data}\}, I) \propto \exp\left(\frac{1}{2} \left[\frac{AB - C^2}{B} \right] (X - X_0)^2\right)$$

1-dim Gaussian pdf

$$\Rightarrow \sigma_x = \sqrt{\frac{-B}{AB - C^2}}$$

$$\text{analogous for } y: \sigma_y = \sqrt{\frac{-A}{AB - C^2}}$$

(incomplete picture)

$$\# \det M = AB - C^2 = \lambda_1 \cdot \lambda_2$$

IF ELLIPSE IS ELONGATED \Rightarrow ONE λ_i is small

$\Rightarrow \det M = 0 \Rightarrow \sigma_x$ AND σ_y is HUGE
(exception $C=0$)

\Rightarrow SO σ_x AND σ_y FAIL TO TELL US THAT THEIR
IS STILL A WELL DETERM. PARAMETER COMB.

SO FAR: error-bar represents $\text{FWHM} \approx 2.35\sigma$

(10)

ALSO: VARIANCE OF POSTERIOR

$$\text{VAR}(X) = \langle (X - \mu)^2 \rangle = \int (X - \mu)^2 \text{prob}(X | \{\text{data}\}, I) dX$$

WITH $\mu = \langle X \rangle$

FOR 1-dim distribution: $\sigma^2 = \langle (X - \mu)^2 \rangle$

STANDARD DEVIATION: Square root of variance
root mean square (r.m.s.)

$$\sigma_x^2 = \langle (X - X_0)^2 \rangle = \iint (X - X_0)^2 \text{prob}(X, Y | \{\text{data}\}, I) dX dY$$

(same for σ_y^2)

ALSO: $\sigma_{xy}^2 = \langle (X - X_0)(Y - Y_0) \rangle$

$$= \iint (X - X_0)(Y - Y_0) \text{prob}(X, Y | \{\text{data}\}, I) dX dY$$

COVARIANCE

CORRELATION BETWEEN INFERRED PARAMETERS

FROM QUADRATIC APPROXIMATION:

$$\sigma_{xy}^2 = \frac{C}{AB - C^2}$$

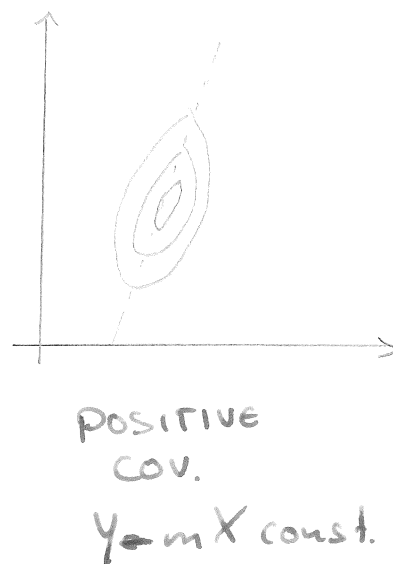
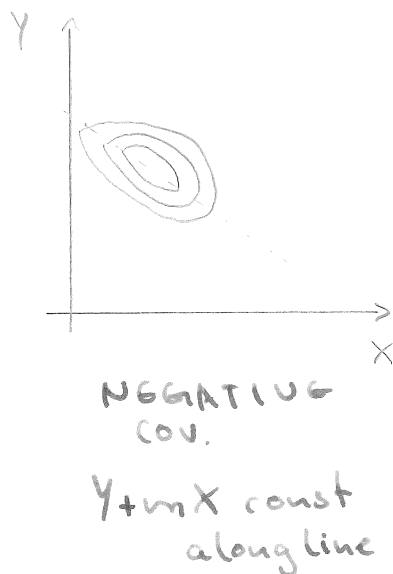
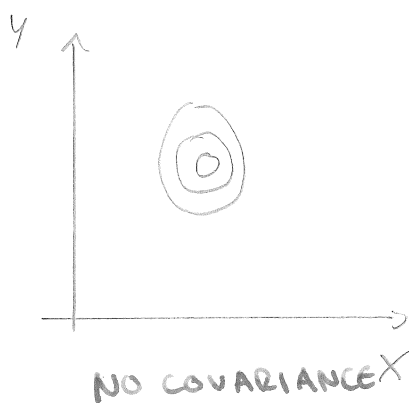
$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix} = \frac{1}{AB - C^2} \begin{pmatrix} -B & C \\ C & -A \end{pmatrix} = - \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1}$$

COVARIANCE MATRIX

FOR $C=0$: $\sigma_{xy}^2 = 0 \Rightarrow$ UNCORRELATED

(11)

\Rightarrow principal direction lie to coordinate axes



EXTREME CASE: $C = \pm \sqrt{AB} \rightarrow$ infinitely wide
 σ_x, σ_y huge

NON

3.2.1 GENERALIZATION OF THE QUADRATIC APPROXIMATION

parameters: $\{X_j\}$

$$\frac{\partial L}{\partial x_i} \Big|_{\vec{x}_0} = 0$$

$$i = 1, 2, \dots, M$$

$$L = L(\vec{x}_0) + \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \frac{\partial^2 L}{\partial x_i \partial x_j} \Big|_{\vec{x}_0} (x_i - x_{0i})(x_j - x_{0j}) + \dots$$

or: $\text{prob}(\vec{x} | \{\text{data}\}, I) \propto \exp \left[\frac{1}{2} (\vec{x} - \vec{x}_0)^T \vec{\nabla} \vec{\nabla} L(\vec{x}_0) (\vec{x} - \vec{x}_0) \right]$

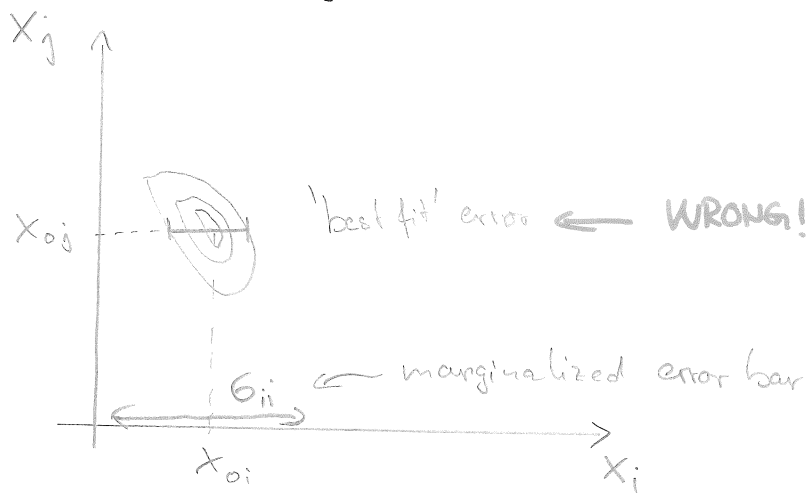
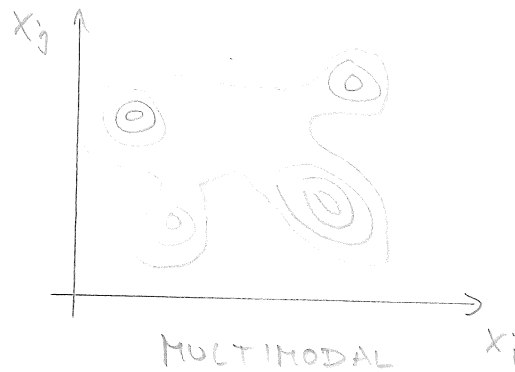
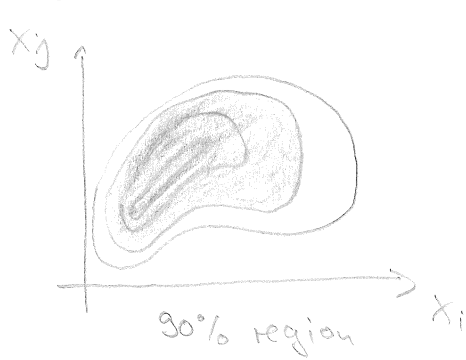
with $\vec{\nabla} \vec{\nabla} L$ symmetric $M \times M$ matrix

MULTIVARIATE GAUSSIAN

MAXIMUM: $\vec{\nabla} L(\vec{x}_0) = 0$

COVARIANCE MATRIX:

$$[\sigma^2]_{ij} = \langle (X_i - X_{oi})(X_j - X_{oj}) \rangle = -[(\hat{\nabla} \hat{\nabla} L)^{-1}]_{ij}$$

3.2.2. ASYMMETRIC AND MULTIMODAL POSTERIOR PDFs

$$\text{prob}(\mu | \{x_k\}, I) = \int_0^{\infty} \text{prob}(\mu, \sigma | \{x_k\}, I) d\sigma$$

Bayes' : $\text{prob}(\mu, \sigma | \{x_k\}, I) \propto \text{prob}(\{x_k\} | \mu, \sigma, I) \times \text{prob}(\mu, \sigma | I)$

$$\text{prob}(\{x_k\} | \mu, \sigma, I) = (\sigma \sqrt{2\pi})^{-N} \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^N (x_k - \mu)^2\right]$$

PRIOR: uniform in μ and $\ln \sigma$ $\nabla \leftarrow$ COMPLETE IGNORANCE

(will be explained later)

use for now:

$$\text{prob}(\mu, \sigma | I) = \begin{cases} \text{constant} & \text{for } \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \text{prob}(\mu | \{x_k\}, I) \propto \int_0^{\infty} t^{N-2} \exp\left[-\frac{t^2}{2} \sum_{k=1}^N (x_k - \mu)^2\right] dt$$

(with $\sigma = 1/t$) use: $\tau = t \sqrt{2 \sum (x_k - \mu)^2}$

$$\Rightarrow \text{prob}(\mu | \{x_k\}, I) \propto \left[\sum_{k=1}^N (x_k - \mu)^2 \right]^{-(N-1)/2}$$

(integral involving only τ can be absorbed in proportionality)

$$\Rightarrow \frac{dL}{d\mu} \Big|_{\mu_0} = \frac{(N-1) \sum (x_k - \mu_0)}{\sum (x_k - \mu_0)^2} = 0$$

$$\Rightarrow \mu_0 = \frac{1}{N} \sum_{k=1}^N x_k$$

$$\frac{d^2 L}{d\mu^2} \Big|_{\mu_0} = - \frac{N(N-1)}{\sum (x_k - \mu_0)^2}$$

$$\Rightarrow \mu = \mu_0 \pm \frac{s}{\sqrt{N}} \quad \text{with} \quad s^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \mu)^2$$

(14)

NOW ESTIMATED
FROM DATA!

3.3.1 THE STUDENT-t and χ^2 DISTRIBUTIONS

ACTUAL pdf of previous analysis:

WRITE:
$$\sum_{k=1}^N (x_k - \mu)^2 \equiv N(\bar{x} - \mu)^2 + V$$

\bar{x} : sample mean and

$$V = \sum_{k=1}^N (x_k - \bar{x})^2$$

$$\Rightarrow \text{prob}(\mu | \{x_k\}, I) \propto [N(\bar{x} - \mu)^2 + V]^{-\frac{N-1}{2}}$$

STUDENT-t DISTRIBUTION

(for $N=3$: CAUCHY DISTRIBUTION)

MAXIMUM: $\mu = \bar{x}$: FWHM $\sim \sqrt{V}$

FOR LARGE N : WINGS disappear \rightarrow Gaussian-like distribution about \bar{x}

optimal value: $\mu_0 = \bar{x}$

errorbar : related to V (see before) ($N \geq 10$)

IF PRIOR $\sim \frac{1}{\sigma}$

(15)

$$\Rightarrow \text{prob}(\mu | \{x_n\}, I) \propto [N(\bar{x} - \mu)^2 + V]^{-N/2}$$

slightly different
to before!

\Rightarrow Student-t distribution with $N-1$ degrees of freedom instead of $N-2$

MAX: \bar{x} ; WIDTH: $\frac{V}{N}$ rather than $\frac{V}{N-1}$
(a little narrower)

$$\text{prob}(\sigma | \{x_n\}, I) = \int_{-\infty}^{+\infty} \text{prob}(\mu, \sigma | \{x_n\}, I) d\mu$$

$$\text{prob}(\sigma | \{x_n\}, I) \propto \sigma^{-N} \exp\left(-\frac{V}{2\sigma^2}\right) \int_{-\infty}^{+\infty} \exp\left[-\frac{N(\bar{x} - \mu)^2}{2\sigma^2}\right] d\mu$$

$$\Rightarrow \text{prob}(\sigma | \{x_n\}, I) \propto \sigma^{N-1} \exp\left(-\frac{V}{2\sigma^2}\right)$$

χ^2 DISTRIBUTION WITH $\chi = \frac{V}{\sigma^2}$

$$\sigma = \sigma_0 \pm \frac{\sigma_0}{\sqrt{2(N-1)}}$$

$$\text{and } \sigma_0 = \sqrt{\frac{V}{N-1}}$$

3.5 Approximations: maximum likelihood and least-squares

\vec{X} : vector of M parameters

\vec{D} : vector of N measured data

Bayes': $\text{prob}(\vec{X} | \vec{D}, I) \propto \text{prob}(\vec{D} | \vec{X}, I) \times \text{prob}(\vec{X} | I)$

large a priori ignorance: $\text{prob}(\vec{X} | I) = \text{const.}$
(uniform, broad)

\Rightarrow broad prior can be absorbed in normalization of posterior

$$\text{prob}(\vec{X} | \vec{D}, I) \propto \text{prob}(\vec{D} | \vec{X}, I)$$

\Rightarrow best estimate \hat{X}_0 of posterior, equivalent to greatest value for probability of observed data

\Rightarrow MAXIMUM LIKELIHOOD ESTIMATE

if data is independent:

$$\text{prob}(\vec{D} | \vec{X}, I) = \prod_{k=1}^N \text{prob}(D_k | \vec{X}, I)$$

from product rule: $\text{prob}(D_k, D_e | \vec{X}, I) = \text{prob}(D_k | D_e, \vec{X}, I) \times \text{prob}(D_e | \vec{X}, I)$

independence: $\text{prob}(D_k | D_e, \vec{X}, I) = \text{prob}(D_k | \vec{X}, I)$

(2)

assume noise associated with measurement is Gaussian

$$\text{prob}(D_k | \vec{X}, I) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left[-\frac{(F_k - D_k)^2}{2\sigma_k^2}\right]$$

I includes: $\{\sigma_k\}$ and ideal noiseless data:

$$F_k = f(\vec{X}, k)$$

$$\Rightarrow \text{prob}(\vec{D} | \vec{X}, I) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

$$\text{with: } \chi^2 = \sum_{k=1}^N \left(\frac{F_k - D_k}{\sigma_k}\right)^2 = \sum_{k=1}^N R_k^2$$

$$R_k = \frac{F_k - D_k}{\sigma_k} : \text{normalized residual}$$

with wide uniform prior:

$$L = \ln \text{prob}(\vec{X} | \vec{D}, I) = \text{constant} - \frac{\chi^2}{2}$$

maximum of posterior $\hat{=}$ minimum of χ^2

\vec{X}_0 is LEAST-SQUARES ESTIMATE

Let's assume functional relationship $f(\vec{x}, k)$ is linear

$$\Rightarrow F_k = \sum_{j=1}^M T_{kj} X_j + C_k$$

(T_{kj}) and (C_k) are independent of \vec{X}

$$\text{matrix notation: } \vec{F} = \mathbf{T} \vec{X} + \vec{C}$$

$$\Rightarrow \frac{\partial L}{\partial X_j} = -\frac{1}{2} \frac{\partial \chi^2}{\partial X_j} = -\sum_{k=1}^N \frac{(F_k - D_k)}{\sigma_k^2} \frac{\partial F_k}{\partial X_j}$$

(3)

$$\text{and } \frac{\partial^2 L}{\partial x_i \partial x_j} = - \sum_{k=1}^N \frac{T_{ki} T_{kj}}{\sigma_k^2}$$

\Rightarrow ALL HIGHER DERIVATIVES ARE ZERO !

\Rightarrow posterior defined by \vec{x}_0 and its covariance matrix related to twice the inverse of $\vec{\nabla} \vec{\nabla} \chi^2$ (Hessian)

$$\langle (x_i - x_{0i})(x_j - x_{0j}) \rangle = - [(\vec{\nabla} \vec{\nabla} \chi^2)^{-1}]_{ij} = 2 [(\vec{\nabla} \vec{\nabla} \chi^2)^{-1}]_{ij}$$

This could also apply to a Poisson distribution, since in the limit of large numbers:

$$\text{prob}(N|D) = \frac{D^N e^{-D}}{N!} \propto \exp\left(-\frac{(N-D)^2}{2D}\right)$$

$$\Rightarrow N \approx D \pm \sqrt{D} \quad \text{and} \quad \chi^2 = \sum_{k=1}^N \frac{(F_k - D_k)^2}{D_k}$$

(all from Gaussian approximation)

$$L_1 - \text{norm} = \sum_{k=1}^N \left| \frac{F_k - D_k}{\sigma_k} \right| \quad \leftarrow \text{will be derived later}$$

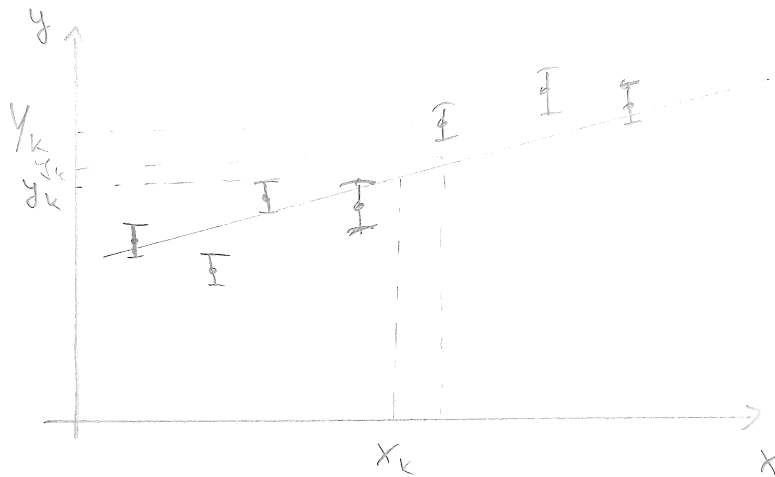
IN GENERAL
WARNING: MAXIMUM LIKELIHOOD ESTIMATE NOT THE SAME AS MOST PROBABLE VALUES OF PARAMETERS !

$$(\text{prob}(A|B) \neq \text{prob}(B|A))$$

3.5.1 FITTING A STRAIGHT LINE

(4)

N data $\{y_k\}$ with errors $\{\sigma_k\}$ at positions $\{x_k\}$



Straight-line model : $y_k = mx_k + c$

Least squares with $F_k = y_k$ and $D_k = y_k$

$$\chi^2 = \sum_{k=1}^N \frac{(mx_k + c - y_k)^2}{\sigma_k^2}$$

$$\frac{\partial \chi^2}{\partial m} = \sum_{k=1}^N \frac{2(mx_k + c - y_k)x_k}{\sigma_k^2}$$

$$\frac{\partial \chi^2}{\partial c} = \sum_{k=1}^N \frac{2(mx_k + c - y_k)}{\sigma_k^2}$$

lets write: $\vec{\nabla} \chi^2 = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} - \begin{pmatrix} p \\ q \end{pmatrix}$

with: $\alpha = \sum w_k x_k^2$

$$\beta = \sum w_k$$

$$\gamma = \sum w_k x_k$$

$$p = \sum w_k x_k y_k$$

$$q = \sum w_k y_k$$

with: $\frac{2}{\sigma_k^2}$

sum $k=1 \dots N$

(m₀, c₀) minimum of χ^2 : $\vec{\nabla} \chi^2 = 0$

⑤

$$\Rightarrow m_0 = \frac{\beta p - \gamma q}{\alpha \beta - \gamma^2} \quad \text{and} \quad c_0 = \frac{\alpha q - \gamma p}{\alpha \beta - \gamma^2}$$

covariance matrix (twice inverse of $\vec{\nabla} \vec{\nabla} \chi^2$)

$$\begin{pmatrix} \sigma_{mm}^2 & \sigma_{mc}^2 \\ \sigma_{mc}^2 & \sigma_{cc}^2 \end{pmatrix} = 2 \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}^{-1} = \frac{2}{\alpha \beta - \gamma^2} \begin{pmatrix} \beta & -\gamma \\ -\gamma & \alpha \end{pmatrix}$$

if error-bar on data is not known assume:

$$\sigma_k = \sigma \quad (\text{same size})$$

$$\Rightarrow w_k = \text{const.} = \frac{2}{\sigma^2}$$

$\Rightarrow m_0$ and c_0 independent of σ

$$\begin{aligned} \text{Uncertainty: } \text{prob}(m, c | \{y_k\}, I) &= \int_0^\infty \text{prob}(m, c, \sigma | \{y_k\}, I) d\sigma \\ &\propto \int_0^\infty \text{prob}(\{y_k\} | m, c, \sigma, I) \times \text{prob}(m, c, \sigma | I) d\sigma \end{aligned}$$

$$\text{prob}(\{y_k\} | m, c, \sigma, I) \propto \sigma^{-N} \exp \left[-\frac{1}{2\sigma^2} \sum_{k=1}^N (m x_k + c - y_k)^2 \right]$$

(keep ALL terms which include σ ; parameter!)

integral over σ as before \Rightarrow Student-t distribution

$$\text{prob}(m, c | \{y_k\}, I) \propto \left[\sum_{k=1}^N (m x_k + c - y_k)^2 \right]^{-(N-1)/2}$$

same best estimate as before, but with σ replaced by

$$S^2 = \frac{1}{N-1} \sum_{k=1}^N (m_0 x_k + c_0 - y_k)^2$$

(still assume x_k precisely known)

3.6 ERROR - PROPAGATION : CHANGING VARIABLES

⑥

EXAMPLE: $X = 10 \pm 3$; $Y = 7 \pm 2$

WHAT CAN WE SAY ABOUT: $X - Y$ or X/Y
or $X^2 + Y^2$

\Rightarrow CHANGING VARIABLES

$\text{prob}(X, Y | I) \rightarrow \text{prob}(Z | I) ?$

with $Z = X - Y$ or $Z = X/Y$

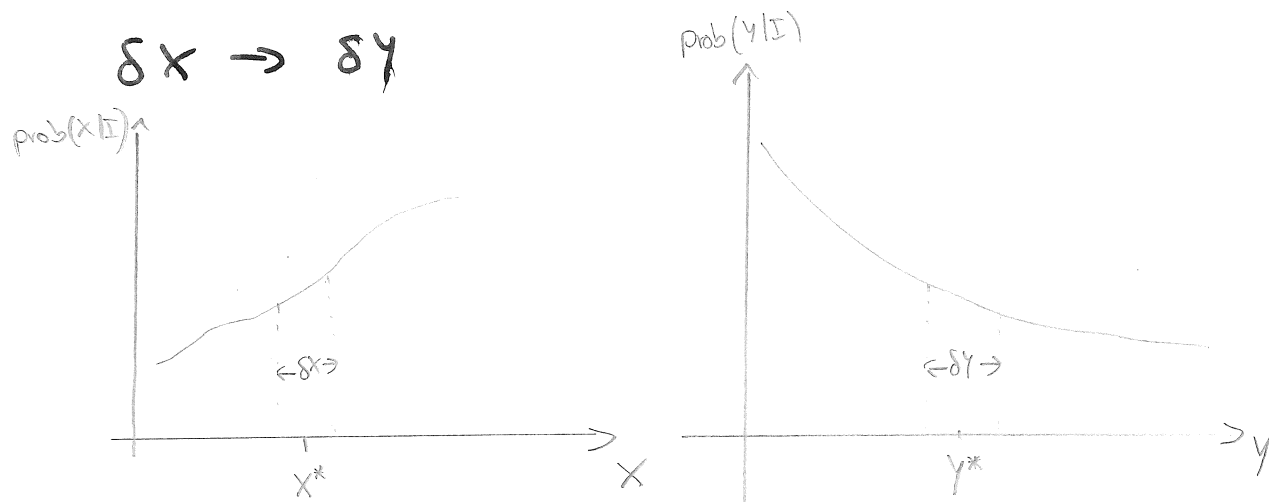
SINGLE VARIABLE: $Y = f(X)$ how is $\text{prob}(X | I)$
related to $\text{prob}(Y | I)$

ASSUME SMALL INTERVAL δX about $X = X^*$

$$\Rightarrow \text{prob} \left(X^* - \frac{\delta X}{2} \leq X \leq X^* + \frac{\delta X}{2} \right) \approx \text{prob}(X = X^* | I) \delta X$$

Y related monotonically to X : $X = X^* \Rightarrow Y = Y^* = f(X^*)$

$\delta X \rightarrow \delta Y$



$$Y^* \pm \frac{\delta Y}{2} \Leftrightarrow X^* \pm \frac{\delta X}{2}$$

$$\text{prob}(X = X^* | I) \delta X = \text{prob}(Y = Y^* | I) \delta Y$$

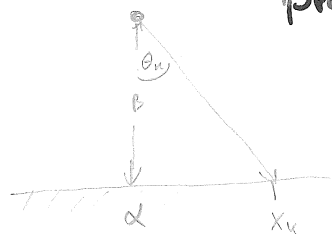
hence: $\text{prob}(x|I) = \text{prob}(y|I) \times \left| \frac{dy}{dx} \right|$

↑ JACOBIAN

EXAMPLE:

LIGHT HOUSE PROBLEM:

$$\text{prob}(\theta|\alpha, \beta, I) = \frac{1}{\pi} \quad \text{with } -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$$



$$\frac{x - \alpha}{\beta} = \tan \theta \Rightarrow \beta \tan \theta = x - \alpha$$

\Rightarrow Differentiate both w.r.t. x

$$\beta \cdot \frac{d \tan \theta}{dx} = \beta \cdot \frac{d \tan \theta}{d\theta} \cdot \frac{d\theta}{dx} = \beta \cdot (1 + \tan^2 \theta) \cdot \frac{d\theta}{dx} = 1$$

$$\left(\tan \theta = \frac{\cos \theta}{\sin \theta} \right)$$

$$\Rightarrow \frac{d\theta}{dx} = \left[\beta (1 + \tan^2 \theta) \right]^{-1} = \left(\beta \left[1 + \frac{(x - \alpha)^2}{\beta^2} \right] \right)^{-1}$$

$$\Rightarrow \text{prob}(x|\alpha, \beta, I) = \text{prob}(\theta|\alpha, \beta, I) \times \left| \frac{d\theta}{dx} \right| = \frac{\beta}{\pi [\beta^2 + (x - \alpha)^2]}$$

GENERALIZATION TO SEVERAL VARIABLES:

$$\text{prob}(\{x_i\}|I) \delta x_1 \delta x_2 \dots \delta x_n = \text{prob}(\{y_i\}|I) \delta^n \text{Vol}(\{y_i\})$$

$\delta^n \text{Vol}(\{y_i\})$: n -dim volume mapped out by small hypercube $\delta x_1 \delta x_2 \dots \delta x_n$

$$\Rightarrow \delta^n \text{Vol}(\{y_i\}) = \left| \frac{\partial (y_1, y_2, \dots, y_n)}{\partial (x_1, x_2, \dots, x_n)} \right| \delta x_1 \delta x_2 \dots \delta x_n$$

$|\dots|$: determinant of $M \times M$ matrix of partial derivatives $\frac{\partial y_i}{\partial x_j}$

$$\text{prob}(\{X_j\} | I) = \text{prob}(\{Y_j\} | I) \times \left| \frac{\partial (Y_1, Y_2, \dots, Y_M)}{\partial (X_1, X_2, \dots, X_M)} \right|$$

ILLUSTRATION: CARTESIAN \rightarrow POLAR COORDINATES

$$x = R \cos \theta$$

$$y = R \sin \theta$$

$$\left| \frac{\partial (x, y)}{\partial (R, \theta)} \right| = \begin{vmatrix} \cos \theta & -R \sin \theta \\ \sin \theta & R \cos \theta \end{vmatrix} = R [\cos^2 \theta + \sin^2 \theta] = R$$

$$\Rightarrow \text{prob}(R, \theta | I) = \text{prob}(x, y | I) \cdot R$$

with Gaussian in (x, y)

$$\Rightarrow \text{prob}(x, y | I) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x^2 + y^2)}{2\sigma^2}\right)$$

$$\text{prob}(R, \theta | I) = \frac{R}{2\pi\sigma^2} \exp\left(-\frac{R^2}{2\sigma^2}\right)$$

obtain pdf for radius: $R = \sqrt{x^2 + y^2}$

$$\text{prob}(R | I) = \int_0^{2\pi} \text{prob}(R, \theta | I) d\theta = \frac{R}{\sigma^2} \exp\left(-\frac{R^2}{2\sigma^2}\right)$$

MULTIDIM. GENERALIZATION:

$$\text{prob}(\vec{D}|\vec{X}, I) \propto \exp\left[-\frac{r_1^2 + r_2^2 + \dots + r_N^2}{2}\right]$$

with $r_k = \frac{F_k - D_k}{\sigma_k}$

$R = \sqrt{\sum r_k^2}$ Square root of χ^2

$\text{prob}(R|\vec{X}, I) dR$ ~ likelihood at that radius \times hypervolume of spherical shell

$$\text{prob}(R|\vec{X}, I) \propto R^{N-1} \exp(-R^2/2)$$

$$\chi^2 = R^2$$

$$d\chi^2 = 2R dR$$

or $\text{prob}(\chi^2|\vec{X}, I) \propto (\chi^2)^{N/2-1} \exp(-\chi^2/2)$

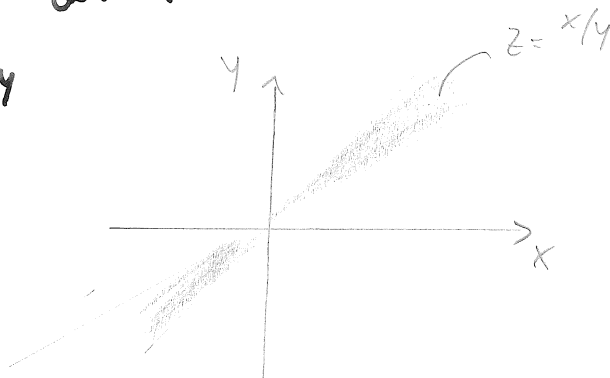
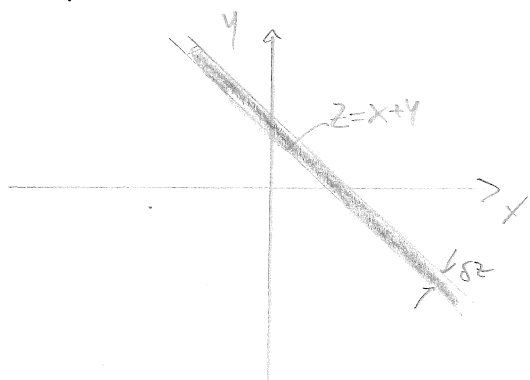
χ^2 distribution with N degrees of freedom

maximum at $N-2$ ($N \geq 2$)

large N : Gaussian with $\chi^2 \approx N \pm \sqrt{2N}$

$Z = X+Y$ or $Z = X/Y$

$\text{prob}(Z|I)$



$$\text{prob}(Z|I) = \iint \text{prob}(Z|X, Y|I) \text{prob}(X, Y|I) dX dY$$

$$= \iint \delta(Z - f(X, Y)) \text{prob}(X, Y|I) dX dY$$

$$Z = X + Y$$

$$\text{prob}(Z|I) = \iint \text{prob}(X, Y|I) \delta(Z - (X + Y)) dx dy$$

assume X and Y are not correlated:

$$X = x_0 \pm \sigma_x \quad Y = y_0 \pm \sigma_y$$

$$\Rightarrow \text{prob}(Z|I) = \int dx \text{prob}(X|I) \int dy \text{prob}(Y|I) \delta(Z - X - Y) dy$$

$$\Rightarrow \text{prob}(Z|I) = \int \text{prob}(X|I) \text{prob}(Y = Z - X|I) dx$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{+\infty} \exp\left[-\frac{(x-x_0)^2}{2\sigma_x^2}\right] \exp\left[-\frac{(Z-x-y_0)^2}{2\sigma_y^2}\right] dx$$

$$\Rightarrow \text{prob}(Z|I) = \frac{1}{\sigma_z \sqrt{2\pi}} \exp\left[-\frac{(Z-z_0)^2}{2\sigma_z^2}\right]$$

with $z_0 = x_0 + y_0$ and $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$

3.6.1. A useful shortcut

$$\langle \delta X^2 \rangle = \sigma_x^2; \quad \langle \delta Y^2 \rangle = \sigma_y^2; \quad \langle \delta X \delta Y \rangle = 0$$

$$\delta Z = \delta X - \delta Y$$

$$\langle \delta Z^2 \rangle = \langle \delta X^2 + \delta Y^2 - 2\delta X \delta Y \rangle = \langle \delta X^2 \rangle + \langle \delta Y^2 \rangle - 2\langle \delta X \delta Y \rangle$$

$$\Rightarrow \sigma_z = \sqrt{\langle \delta Z^2 \rangle} = \sqrt{\sigma_x^2 + \sigma_y^2}$$

$$z = x/y$$

$$\delta z = \frac{y \delta x - x \delta y}{y^2} \Rightarrow \frac{\delta z}{z} = \frac{\delta x}{x} - \frac{\delta y}{y}$$

$$\Rightarrow \frac{\langle \delta z^2 \rangle}{z_0^2} = \frac{\langle \delta x^2 \rangle}{x_0^2} + \frac{\langle \delta y^2 \rangle}{y_0^2} - \frac{2 \langle \delta x \delta y \rangle}{x_0 y_0}$$

$$\Rightarrow \frac{\sigma_z}{z_0} = \sqrt{\left(\frac{\sigma_x}{x_0}\right)^2 + \left(\frac{\sigma_y}{y_0}\right)^2}$$

Lecture V

①

3.6.2. Taking the square root of a number

- example: peak of Gaussian amplitude is related to complex structure factor

$$A = |F|^2$$

write: $A = f^2$ with $f = |F|$

find: best estimate f_0 and reliability σ_f

of course: $f_0 = \sqrt{A_0}$

and: $\langle \delta A^2 \rangle = 4 f_0^2 \langle \delta f^2 \rangle = 4 A_0 \langle \delta f^2 \rangle$

with: $\sigma_A^2 = \langle \delta A^2 \rangle$

$$\Rightarrow f = \sqrt{A_0} \pm \frac{\sigma_A}{2\sqrt{A_0}}$$

Does not work for negative A_0 !

(yes this can happen) \Leftarrow TWO MISTAKES HERE!

1.) FAILURE TO DISTINGUISH posterior and likelihood

least-square fit: $\text{prob}(\{\text{data}\} | A, I) \propto \exp\left(-\frac{(A - A_0)^2}{2\sigma_A^2}\right)$

however we need:

$$\text{prob}(A | \{\text{data}\}, I) \propto \text{prob}(\{\text{data}\} | A, I) \times \text{prob}(A | I)$$

we know: ~~ap~~ amplitude must be positive

(2)

$$\text{naïve choice : } \text{prob}(A|I) = \begin{cases} \text{const} & \text{for } A \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

\Rightarrow best estimate always positive

true even if A_0 which gives closest agreement with data is negative

posterior pdf is severely truncated

\Rightarrow procedure of error-propagation **FLAWED!**
(relies on expansion around maximum)

\uparrow SECOND
MISTAKE!

\hookrightarrow GO BACK TO PROPER CHANGE OF VARIABLES!

$$\text{prob}(f|\{\text{data}\}, I) = \text{prob}(A|\{\text{data}\}, I) \times \left| \frac{dA}{df} \right|$$

$$\text{Jacobian: } \left| \frac{dA}{df} \right| = 2f \quad \text{with } f = |F| \geq 0$$

$$\Rightarrow \text{prob}(f|\{\text{data}\}, I) \propto f \exp \left[-\frac{(f^2 - A_0)^2}{2\sigma_A^2} \right] \quad \text{with } f \geq 0$$

and zero otherwise

Gaussian approximation:

first and second derivatives of $L = \ln[\text{prob}(f|\{\text{data}\}, I)]$

$$\Rightarrow 2f_0^2 = A_0 + (A_0^2 + 2\sigma_A^2)^{1/2}$$

$$\text{and: } \sigma_f^{-2} = \frac{1}{f_0^2} + \frac{2(3f_0^2 - A_0)}{\sigma_A^2}$$

reduces to "standard" error-propagation result
in the case $A_0 \gg \sigma_A$

(3)

Lecture 5.1.R

$$A = 9 \pm 1$$

$$A = 1 \pm 9$$

$$A = -20 \pm 9$$

RULES OF PROBABILITY THEORY ARE SAFER TO USE
THAN 'COOK-BOOK' APPROACHES

4 MODEL SELECTION

QUESTIONS BEYOND PARAMETER FITTING

for example:

- is a quadratic or cubic function more appropriate than a linear function?
- is a Gaussian or Cauchy distribution better to describe a peak?

Mr A has a theory; Mr B also has a theory, but with one adjustable parameter λ . Whose theory should we prefer on the basis of data D?

Mr A : y vs x
 $y = 0$

Mr B : $y = a$

Mr C : $y = a + bx$

relative merit of two theories:

$$\text{posterior ratio} = \frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)}$$

greater than one \Rightarrow theory A

less than one \Rightarrow theory B

≈ 1

\Rightarrow can not make judgement

RATIO
OF
PRIORS

Bayes' theorem:
$$\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(D|A, I)}{\text{prob}(D|B, I)} \times \frac{\text{prob}(A, I)}{\text{prob}(B, I)}$$

(prob(D|I) drops out)

COULD TAKE RATIO OF PRIORS UNITY

(NOT ALWAYS THE CASE)

MR B has a free parameter λ , however:

$$\text{prob}(D|B, I) = \int \text{prob}(D, \lambda|B, I) d\lambda$$

$$= \int \text{prob}(D|\lambda, B, I) \text{prob}(\lambda|B, I) d\lambda$$

\uparrow
LIKELIHOOD

\uparrow
PRIOR ON λ

ASSUME: λ must lie between λ_{\min} and λ_{\max}

(5)

$$\text{prob}(\lambda | B, I) = \frac{1}{\lambda_{\max} - \lambda_{\min}} \quad \text{for } \lambda_{\min} \leq \lambda \leq \lambda_{\max}$$

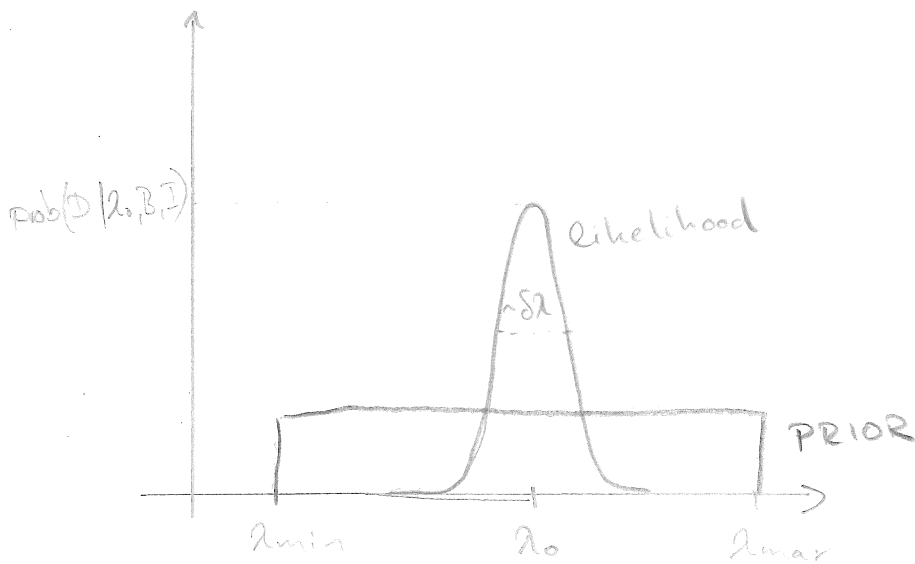
zero otherwise

assume: λ_0 : closest agreement with measurement!

$\text{prob}(D | \lambda_0, B, I)$ maximum of likelihood

and $\lambda_0 \pm \delta\lambda$ Gaussian approx (assumption)

$$\text{prob}(D | \lambda, B, I) = \text{prob}(D | \lambda_0, B, I) \times \exp\left[-\frac{(\lambda - \lambda_0)^2}{2\delta\lambda^2}\right]$$



$$\text{prob}(D | B, I) = \frac{1}{\lambda_{\max} - \lambda_{\min}} \int_{\lambda_{\min}}^{\lambda_{\max}} \text{prob}(D | \lambda, B, I) d\lambda$$

(assume no truncation of Gaussian by prior)

$$\Rightarrow \text{prob}(D | B, I) = \frac{\text{prob}(D | \lambda_0, B, I) \times \delta\lambda \sqrt{2\pi}}{\lambda_{\max} - \lambda_{\min}}$$

$$\Rightarrow \frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(A|I)}{\text{prob}(B|I)} \times \underbrace{\frac{\text{prob}(D|A, I)}{\text{prob}(D|\lambda_0, B, I)}}_{\substack{\text{GOODNESS} \\ \text{OF FIT}}} \times \underbrace{\frac{\lambda_{\max} - \lambda_{\min}}{\delta\lambda \sqrt{2\pi}}}_{\substack{\text{Occam's} \\ \text{razor}}} \quad (6)$$

↑
UNITY: FAIRNESS

HOW WELL DO BEST PREDICTIONS AGREE WITH DATA:

⇒ ALWAYS FAVOURS Mr B (with one parameter)

IF THIS WOULD BE THE ONLY TERM

MORE COMPLICATED THEORY ALWAYS FAVOURED!

prior range $\lambda_{\max} - \lambda_{\min}$ generally much larger than $\pm \delta\lambda \Rightarrow$ **FINAL TERM** penalizes B

Ochham factor \longleftrightarrow Ochham's razor:

"it is vain to do with more what can be done with fewer"

Jeffreys (1939): infinite penalty if any new parameter was allowed to go $\pm \infty$
(but if we really want this we have a theory with no predictable power)

usually we can motivate tighter physical priors.

Assume now $\pi_r A$ also has an adjustable parameter μ

$$\Rightarrow \frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(A|I)}{\text{prob}(B|I)} \times \frac{\text{prob}(D|\mu_0, A, I)}{\text{prob}(D|\lambda_0, B, I)} \times \frac{\delta\mu (\lambda_{\max} - \lambda_{\min})}{\delta\lambda (\lambda_{\max} - \lambda_{\min})}$$

if equal weight is given to A and B
and same prior range assigned:

$$\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} \approx \frac{\text{prob}(D|\mu_0, A, I)}{\text{prob}(D|\lambda_0, B, I)} \times \frac{\delta\mu}{\delta\lambda}$$

if both give similar likelihoods; shape with larger
error-bar is preferred! \Leftarrow MORE PARAMETER VALUES ARE
CONSISTENT WITH DATA!

ASSUME CASE FOR SAME PHYSICAL THEORY BUT DIFFERENT
PRIORS

$$\Rightarrow \frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\lambda_{\max} - \lambda_{\min}}{\mu_{\max} - \mu_{\min}}$$

\Rightarrow prefer theorist with narrower prior range
(more insight!)

4.1.1. COMPARISON TO PARAMETER ESTIMATION

Bayes' theorem

$$\text{prob}(\lambda|D, B, I) = \frac{\text{prob}(D|\lambda, B, I) \times \text{prob}(\lambda|B, I)}{\underbrace{\text{prob}(D|B, I)}_{\text{usually omitted}}}$$

\nwarrow
plays crucial rôle in model comparison

EVIDENCE

parameter estimation: location of maximum of posterior

model selection: average value of posterior

→ if prior is uniform (flat) and wide it does not matter for parameter estimation, however it matters for the range over which the average is taken

↳ Ockham's razor or goodness-of-fit

more suitable prior sometimes:

$$\text{prob}(\lambda | \mathcal{B}, I) = \frac{e^{-\lambda/b}}{b}$$

4.1.2 Hypothesis Testing

↳ conventional ("frequentist") way to deal with model selection

hypothesis H_1 : shape of signal peak is Gaussian

$$\text{prob}(H_1 | \mathcal{D}, I) = \frac{\text{prob}(\mathcal{D} | H_1, I) \times \text{prob}(H_1 | I)}{\text{prob}(\mathcal{D} | I)}$$

odds-ratio:
$$\frac{\text{prob}(H_1 | \mathcal{D}, I)}{\text{prob}(H_2 | \mathcal{D}, I)} = \frac{\text{prob}(\mathcal{D} | H_1, I)}{\text{prob}(\mathcal{D} | H_2, I)} \times \frac{\text{prob}(H_1 | I)}{\text{prob}(H_2 | I)}$$

assume: $H_2 = \bar{H}_1$ (H_1 is false)

$$\text{prob}(\mathcal{D} | I) = \text{prob}(\mathcal{D} | H_1, I) \times \text{prob}(H_1 | I) + \text{prob}(\mathcal{D} | \bar{H}_1, I) \times \text{prob}(\bar{H}_1 | I)$$

priors related by sum rule:
$$\text{prob}(H_1 | I) + \text{prob}(\bar{H}_1 | I) = 1$$

Problem: $\text{prob}(\mathcal{D} | \bar{H}_1, I)$

(e.g. if the shape is not Gaussian, what is it then?)

historically: assess significance of mismatch

$\rightarrow \chi^2$ - statistics used

this is only true for Gaussian noise: expected value

$\chi^2 \sim N$ (number of measurements)

would not expect deviations from a few $\sim \sqrt{N}$

HOWEVER THESE STATEMENTS ARE ABOUT THE

LIKELIHOOD AND NOT THE POSTERIOR!

PRIOR CAN CHANGE THIS ASSIGNMENT

4.2. EXAMPLE: HOW MANY LINES ARE THERE?

QUESTION: HOW MANY SIGNAL PEAKS THERE IS MOST EVIDENCE? ^{FOR}

infer amplitude and position

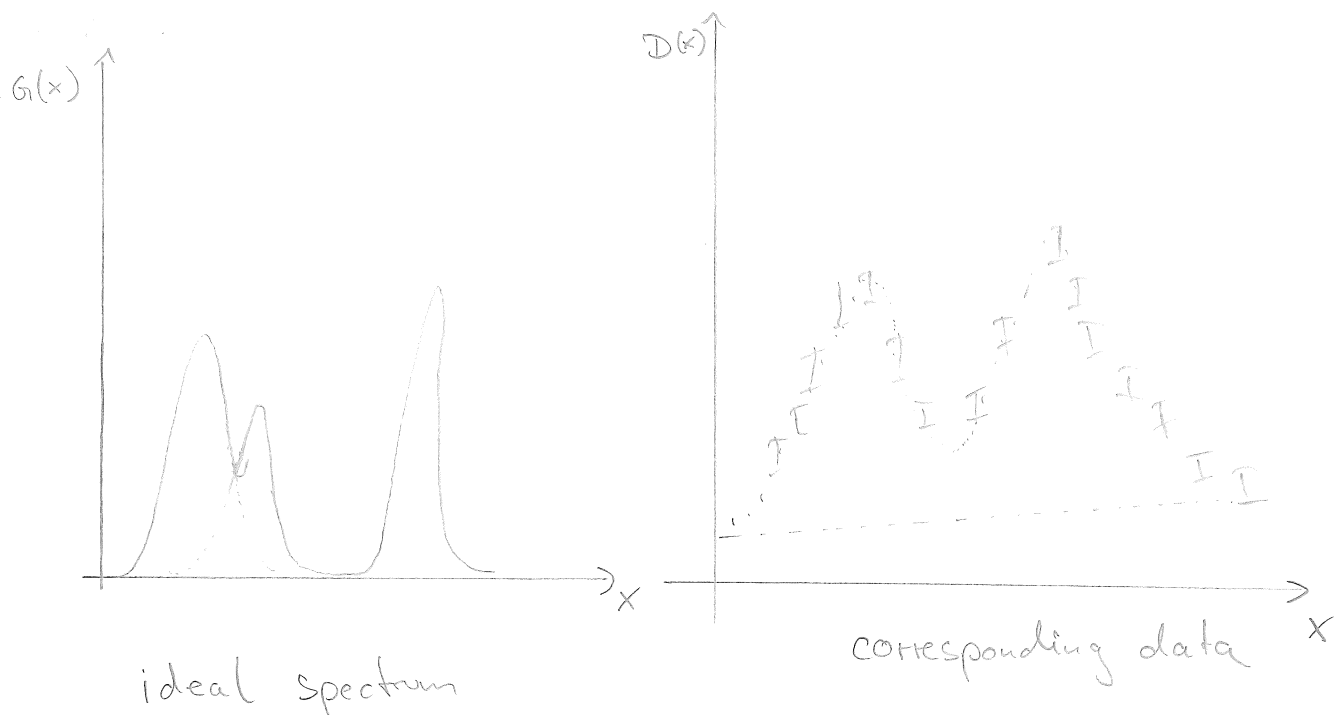
assume: shape of signal peaks known

ideal spectrum: $G(x) = \sum_{j=1}^M A_j f(x, x_j)$

A_j : magnitude of j -th line

x_j : location

with Gaussian: $f(x, x_j) = \exp \left[-\frac{(x - x_j)^2}{2W^2} \right]$



include background $B(x)$

ideal data $\{F_k\}$

$$F_k = \int G(x) R(x_k - x) dx + B(x_k)$$

↪ assume: shape of resolution function does not vary with position!

blurring of Gaussian

assume: data $\{D_k\}$ independent, additive Gaussian noise $\{G_k\}$

⇒ least-squares likelihood:

$$\text{prob}(\{D_k\} | \{A_{ij}, x_{ij}\}, M, I) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

χ^2 : least-squares

normalization: product of $\frac{1}{G_k \sqrt{2\pi}}$

estimate number of signal peaks:

$$\text{prob}(M|\{D_n\}, I) = \frac{\text{prob}(\{D_n\}|M, I) \times \text{prob}(M|I)}{\text{prob}(\{D_n\}|I)}$$

assign uniform prior for $M=1, 2, \dots$ up to a few (20 or so)

$$\Rightarrow \text{prob}(M|\{D_n\}, I) \propto \text{prob}(\{D_n\}|M, I)$$

(normalization calculated later)

$$\Rightarrow \sum \text{prob}(M|\{D_n\}, I) = 1$$

~~As in case of Mr B in the previous section~~

MARGINAL INTEGRAL OVER PRODUCT OF PRIOR
AND LIKELIHOOD FUNCTION:

$$\text{prob}(\{D_n\}|M, I) = \iint \dots \int \text{prob}(\{D_n\}, \{A_j, x_j\}|M, I) d^M A_j d^M x_j$$

where

$$\text{prob}(\{D_n\}, \{A_j, x_j\}|M, I) = \text{prob}(\{D_n\}|\{A_j, x_j\}, M, I) \text{prob}(\{A_j, x_j\}|M, I)$$

again, simple uniform pdf : $x_{\min} \leq x_j \leq x_{\max}$
 $0 \leq A_j \leq A_{\max}$

$$\Rightarrow \text{prob}(\{A_j, x_j\}|M, I) = [(x_{\max} - x_{\min}) A_{\max}]^{-M}$$

$$\text{prob}(M|\{D_n\}, I) \propto [(x_{\max} - x_{\min}) A_{\max}]^{-M} \underbrace{\iint \dots \int \exp\left(-\frac{x^2}{2}\right) d^M A_j d^M x_j}_{\text{over defined prior region}}$$

THIS INTEGRAL IS BEST DONE NUMERICALLY
(MONTE-CARLO)

CRUDE SIMPLIFYING APPROXIMATIONS:

TAYLOR: $\chi^2 \approx \chi^2_{\min} + \frac{1}{2} (\vec{x} - \vec{x}_0)^T \vec{\nabla} \vec{\nabla} \chi^2(\vec{x}_0) (\vec{x} - \vec{x}_0) + \dots$

$$\vec{x}_0 = \{A_{0j}, x_{0j}\}$$

INTEGRAL TO SOLVE:

$$\begin{aligned} \iint \dots \int \exp \left[-\frac{1}{2} (\vec{x} - \vec{x}_0)^T \vec{\nabla} \vec{\nabla} \chi^2(\vec{x}_0) (\vec{x} - \vec{x}_0) \right] d^N x_j \\ = \frac{(4\pi)^N}{\sqrt{\det(\vec{\nabla} \vec{\nabla} \chi^2)}} \end{aligned}$$

$\det(\vec{\nabla} \vec{\nabla} \chi^2)$ determinant of Hessian matrix, evaluated \vec{x}_0
 since labelling is arbitrary there are $M!$ equivalent maxima in the likelihood (numbers of permutations)

$$\Rightarrow \text{prob}(M | \{D_N\}, I) \propto \frac{M! (4\pi)^N}{[(x_{\max} - x_{\min}) A_{\max}]^N \sqrt{\det(\vec{\nabla} \vec{\nabla} \chi^2)}} \exp\left(-\frac{\chi^2_{\min}}{2}\right)$$

ALGORITHM:

NEED TO FIND χ^2_{\min} and determinant $\vec{\nabla} \vec{\nabla} \chi^2$
 (ingredient, standard least squares!)
 nls in R?

* NOT SURE WHAT TO IMPLEMENT IN R?

MAYBE SIMPLE CASE?

Generate linear data and fit with constant, line and parabola. Calculate evidence and see what happens.

4.3. Other Examples: Means, variance, dating and so on

PROBLEM OF CLASSIFICATION:

(Archeology): Suppose that two sites yield N_1 and N_2 measurements: \vec{D}_1 and \vec{D}_2

↑
evidence for
evolution
(change in trait size)

TWO HYPOTHESIS (MODELS):

A: THERE IS NO DIFFERENCE BETWEEN THE TWO DATA SETS: characterized by same unknown mean and std. σ

B: THERE IS A DIFFERENCE: μ_1, μ_2
 σ_1, σ_2

Need to evaluate evidence:

$$\text{prob}(\vec{D}_1, \vec{D}_2 | A, I) \quad \text{and} \quad \text{prob}(\vec{D}_1, \vec{D}_2 | B, I)$$

$$\text{prob}(\vec{D}_1, \vec{D}_2 | A, I) = \iint \text{prob}(\vec{D}_1, \vec{D}_2 | \mu, \sigma, A, I) \text{prob}(\mu, \sigma | A, I) d\mu d\sigma$$

prior for A: uniform: $\mu_{\min} \leq \mu \leq \mu_{\max}$
 $0 \leq \sigma \leq \sigma_{\max}$

$$\Rightarrow \text{prob}(\mu, \sigma | A, I) = \frac{1}{(\mu_{\max} - \mu_{\min}) \sigma_{\max}}$$

ASSIGN A GAUSSIAN LIKELIHOOD FCT.

(why? see next chapter)

SINGLE DATA SET: $N = N_1 + N_2$

$$\Rightarrow \text{prob}(\vec{D}_1, \vec{D}_2 | \mu, \sigma, A, I) = (\sigma \sqrt{2\pi})^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^N (x_k - \mu)^2 \right\}$$

\Rightarrow CALCULATE INTEGRAL NUMERICALLY

APPROXIMATION: TAYLOR EXPAND ^{Likelihood} logarithm of ~~posterior~~:

$$L = L(\mu_0, \sigma_0) - \frac{1}{2} (\mu - \mu_0 \quad \sigma - \sigma_0) \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} \begin{pmatrix} \mu - \mu_0 \\ \sigma - \sigma_0 \end{pmatrix} + \dots$$

with $L = \ln [\text{prob}(\vec{D}_1, \vec{D}_2 | \mu, \sigma, A, I)]$

maximum at (μ_0, σ_0)

$$\frac{\partial L}{\partial \mu} = 0$$

$$\frac{\partial L}{\partial \sigma} = 0$$

$$\mu_0 = \frac{1}{N} \sum_{k=1}^N x_k$$

$$; \quad \sigma_0^2 = \frac{1}{N} \sum_{k=1}^N (x_k - \mu_0)^2$$

second derivatives:

$$\alpha = \frac{N}{\sigma_0^2} \quad \text{and} \quad \beta = 2 \frac{N}{\sigma_0^3} \quad ; \quad \gamma = 0$$

exponentiating gives for 2nd order term:

$$\int_{\mu_{\min}}^{\mu_{\max}} \int_0^{\sigma_{\max}} \exp \left(-\frac{1}{2} \left[\alpha (\mu - \mu_0)^2 + \beta (\sigma - \sigma_0)^2 \right] \right) d\mu d\sigma \approx \frac{2\pi}{\sqrt{\alpha\beta}}$$

THIS HAS TO BE MULTIPLIED BY

(2)

$$\frac{\exp[L(\mu_0, \sigma_0)]}{(\mu_{\max} - \mu_{\min}) \sigma_{\max}} \quad \text{to obtain posterior}$$

insert relations for α, β, μ_0 and σ_0

$$\Rightarrow \text{prob}(\vec{D}_1, \vec{D}_2 | A, I) \approx \frac{(\sigma_0 \sqrt{2\pi})^{2-N} \exp(-N/2)}{(\mu_{\max} - \mu_{\min}) \sigma_{\max} N \sqrt{2}}$$

FOR OTHER MODEL B

$$\text{prob}(\vec{D}_1, \vec{D}_2 | B, I) = \text{prob}(\vec{D}_1 | B, I) \times \text{prob}(\vec{D}_2 | B, I)$$

$$\text{prob}(\vec{D}_j | B, I) = \iint \text{prob}(\vec{D}_j | \mu_j, \sigma_j, B, I) \text{prob}(\mu_j, \sigma_j | B, I) d\mu_j d\sigma_j$$

with $j = 1, 2$: Set two priors identical to previous

$$\text{same as before: } \text{prob}(\vec{D}_j | B, I) \approx \frac{(\sigma_{0j} \sqrt{2\pi})^{2-N_j} \exp(-N_j/2)}{(\mu_{\max} - \mu_{\min}) \sigma_{\max} N_j \sqrt{2}}$$

ratio of ^{posterior} evidence:

$$\frac{\text{prob}(\vec{D}_1, \vec{D}_2 | A, I)}{\text{prob}(\vec{D}_1, \vec{D}_2 | B, I)} \approx \frac{(\mu_{\max} - \mu_{\min}) \sigma_{\max}}{\pi \sqrt{2}}$$

$$\times \frac{N_1 N_2 (\sigma_0)^{2-N}}{N (\sigma_{01})^{2-N_1} (\sigma_{02})^{2-N_2}}$$

From Bayes' : ratio of evidence

$$\text{ratio of posterior} \times \underbrace{\frac{\text{prob}(A|I)}{\text{prob}(B|I)}}_1$$

Generalization to M data sets:

$$\frac{\text{prob}(\{D_j\} | A, I)}{\text{prob}(\{D_j\} | B, I)} \approx \left[\frac{(\mu_{\max} - \mu_{\min}) \sigma_{\max}}{\pi \sqrt{2}} \right]^{M-1} \times \frac{(\sigma_0)^{2-N}}{N} \times \prod_{j=1}^M \frac{N_j}{(\sigma_{0j})^{2-N_j}}$$

Example: ~~lifetime~~ of ~~life time~~ of lightbulbs
from two manufactures

posterior probability of hypothesis $\mu_1 > \mu_2$

$$\text{prob}(\mu_1 > \mu_2 | \vec{D}_1, \vec{D}_2, I) = \int_0^\infty d\mu_1 \int_0^{\mu_1} d\mu_2 \text{prob}(\mu_1, \mu_2 | \vec{D}_1, \vec{D}_2, I)$$

if this prob close to 1 manufact. 1 would be better, close to zero manufact. 2

close to 0.5 no strong preference

mean of both manufactures independent

$$\text{prob}(\mu_1, \mu_2 | \vec{D}_1, \vec{D}_2, I) = \underbrace{\text{prob}(\mu_1 | \vec{D}_1, I)}_{\text{MARGINAL INTEGRAL OVER VARIANCE } \sigma_1} \times \text{prob}(\mu_2 | \vec{D}_2, I)$$

MARGINAL INTEGRAL
OVER VARIANCE σ_1

\Rightarrow Student - t
DISTRIBUTION

$$\Rightarrow \text{prob}(\mu_1, \mu_2 | \vec{D}_1, \vec{D}_2, I) \approx \frac{\sqrt{N_1 N_2}}{2\pi S_1 S_2} \exp \left[-\frac{1}{2} \left[\frac{N_1 (\mu_1 - \mu_{01})^2}{S_1^2} + \frac{N_2 (\mu_2 - \mu_{02})^2}{S_2^2} \right] \right]$$

(5)

$\mu_{01}, \mu_{02}, S_1, S_2$ are obtained by the corresponding sums

introduce new variable $Z = \mu_1 - \mu_2$

$$\Rightarrow \text{prob}(\mu_1 > \mu_2 \mid \vec{D}_1, \vec{D}_2, \vec{I}) \approx \frac{1}{S_2 \sqrt{2\pi}} \int_0^{\infty} \exp\left[-\frac{(Z - z_0)^2}{2S_2^2}\right] dZ$$

= *

with $z_0 = \mu_{01} - \mu_{02}$

$$S_2^2 = \frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}$$

performing the integration : $* = \frac{1}{2} + \frac{1}{2} \text{erf}\left(\frac{z_0}{\sqrt{2} S_2}\right)$

$$z_0 = S_2 \Rightarrow \frac{1}{2} + \frac{1}{2} 0.68 = 0.84$$

$$z_0 = 2S_2 \Rightarrow 0.98$$

if z_0 is negative 1 - these numbers

5 ASSIGNING PROBABILITIES

⑥

- HOW DO WE DECIDE WHICH DISTRIBUTION TO ASSIGN TO THE LIKELIHOOD?

5.1 IGNORANCE: INDIFFERENCE AND TRANSFORMATION GROUPS

Bernoulli (1713): PRINCIPLE OF INSUFFICIENT REASON.

Keynes (1921): PRINCIPLE OF INDIFFERENCE

DEVISE A SET OF BASIC, MUTUALLY EXCLUDING
POSSIBILITIES \leftarrow ASSIGN SAME PROBABILITY
TO EACH

EXAMPLE: ORDINARY DICE

POTENTIAL OUTCOMES: $X_i \equiv$ the face on top has
 i dots

$i = 1, 2, \dots, 6$

Bernoulli: $\text{prob}(X_i | I) = 1/6$

CAN THIS BE JUSTIFIED IN A MORE FUNDAMENTAL
WAY?

CALL SIX FACES A, B, C, D, E, F

$\Rightarrow \text{prob}(A | I), \text{prob}(B | I), \dots, \text{prob}(F | I)$
 $\downarrow \quad \quad \quad \downarrow \quad \quad \quad \downarrow$
 $X_1 \quad \quad \quad X_2 \quad \quad \quad X_6$

\Rightarrow WOULD RENAMING CHANGE ANYTHING?
NO (IF GROSS IGNORANCE WAS APPLIED)

\Rightarrow JAYNES (1978) "DESIDERATUM OF CONSISTENCY"

EXAMPLE: COLOURED BALLS DRAWN RANDOMLY

CONTENTS: W: white balls ; R: red balls

principle of indifference: $\text{prob}(j | I) = \frac{1}{R+W}$

for proposition that any particular ball (index j) will be drawn

from marginalization:

$$\text{prob}(\text{red} | I) = \sum_{j=1}^{R+W} \text{prob}(\text{red}, j | I) = \frac{1}{R+W} \underbrace{\sum_{j=1}^{R+W} \text{prob}(\text{red} | j, I)}_{R: \text{ since } R \text{ red balls}}$$

$$= \frac{R}{R+W}$$

SAMPLING WITH REPLACEMENT:

PROBABILITY that N trials result in r red balls

$$\begin{aligned} \text{prob}(r | N, I) &= \sum_k \text{prob}(r, S_k | N, I) \\ &= \sum_k \text{prob}(r | S_k, N, I) \text{prob}(S_k | N, I) \end{aligned}$$

where the sum is over 2^N possible sequences of red-white outcomes $\{S_k\}$

$$\text{prob}(r | S_k, N, I) = \begin{cases} 1 & \text{if } S_k \text{ contains exactly } r \text{ red balls} \\ 0 & \text{otherwise} \end{cases}$$

\Rightarrow only need to consider S_k with exactly r red balls for $\text{prob}(S_k | N, I)$

(8)

replacement \Rightarrow result of one draw, does not influence next draw

probability of drawing any particular sequence only depends on total number of red and white balls

$$\text{prob}(S_n | N, I) = [\text{prob}(\text{red} | I)]^r \times [\text{prob}(\text{white} | I)]^{N-r}$$

$$= \left[\frac{R}{R+W} \right]^r \times \left[\frac{W}{R+W} \right]^{N-r}$$

$$= \frac{R^r W^{N-r}}{(R+W)^N}$$

MUST BE MULTIPLIED BY NUMBER OF POSSIBLE SEQUENCES WHICH CONTAIN EXACTLY r BALLS IN N -draws

EASIER:

1) IN HOW MANY WAYS CAN N DIFFERENT OBJECTS BE ARRANGED ON A STRAIGHT LINE

N choices for first item

$(N-1)$ " " 2nd item

$(N-2)$

$(N-3)$

\Rightarrow TOTAL NUMBER OF PERMUTATIONS: $N \times (N-1) \times \dots \times 3 \times 2 \times 1 = N!$

2) IN HOW MANY WAYS CAN WE PICK M OBJECTS FROM N DIFFERENT ONES:

$$N \times (N-1) \times (N-2) \times \dots \times (N-M+2) \times (N-M+1) \equiv {}^N P_M$$

$${}^N P_M = \frac{N!}{(N-M)!}$$

$$(M=N \Rightarrow N!)$$

$$0! = 1! = 1$$

3) IF WE ARE NOT INTERESTED IN ORDER OF
THEM OBJECTS: DIVIDE BY NUMBER M OBJECTS
CAN BE ORDERED

(9)

$${}^N C_M = \frac{N!}{M!(N-M)!}$$

KNOWN FROM BINOMIAL EXPANSION:

$$(a+b)^N = \sum_{j=0}^N \frac{N!}{j!(N-j)!} a^j b^{N-j}$$

if ~~$a+b=1 \Rightarrow \sum_{j=0}^N \frac{N!}{j!(N-j)!} = 1$~~

$$a=b=1 \Rightarrow \sum_{m=0}^n {}^n C_m = 2^n$$

problem for red balls: select r integers out of N
where order is irrelevant: ${}^N C_r$

$$\Rightarrow \text{prob}(r | N, I) = \frac{N!}{r!(N-r)!} \times \frac{R^r W^{N-r}}{(R+W)^N}$$

check normalization:

$$\sum_{r=0}^N \text{prob}(r | N, I) = \sum_{r=0}^N \frac{N!}{r!(N-r)!} p^r q^{N-r} = (p+q)^N = 1^N = 1$$

$p+q=1$ since $p = \frac{R}{R+W}$ $q = \frac{W}{R+W}$

$$\left\langle \frac{r}{N} \right\rangle = \sum_{r=0}^N \frac{r}{N} \text{prob}(r | N, I) = \sum_{r=1}^N \frac{(N-1)!}{(r-1)!(N-r)!} p^r q^{N-r}$$

contribution of $r=0$ is zero

$$j = r-1$$

$$\begin{aligned}\left\langle \frac{r}{N} \right\rangle &= p \sum_{j=0}^{N-1} \frac{(N-1)!}{j!(N-1-j)!} p^j q^{N-1-j} \\ &= p (p+q)^{N-1} = p = \frac{R+r}{R+W}\end{aligned}$$

$$\text{variance : } \left\langle \left(\frac{r}{N} - p \right)^2 \right\rangle = \frac{pq}{N} \rightarrow 0 \text{ for } N \rightarrow \infty$$

Bernoulli's theorem for large numbers

$$\lim_{N \rightarrow \infty} \left(\frac{r}{N} \right) = \text{prob}(\text{red} | I)$$

5.1.1 THE BINOMIAL DISTRIBUTION

2 possible outcomes: success and failure

$$\text{prob}(\text{success} | I) = p \quad \text{prob}(\text{failure} | I) = q = 1-p$$

r successes in N trials

$$\text{prob}(r | N, I) = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r}$$

$$r = 0, 1, 2, \dots, N$$

BINOMIAL DISTRIBUTION

$$\langle r \rangle = Np \quad \text{and} \quad \langle (r - Np)^2 \rangle = Np(1-p)$$

GENERALIZATION TO CONTINUOUS PARAMETERS

$$\text{prob}(X=x|I) = \lim_{\delta x \rightarrow 0} \text{prob}(x \leq X < x + \delta x | I)$$

consistency : pdf should change little if offset x_0

$$\text{prob}(X|I) dX \approx \text{prob}(X+x_0|I) \underbrace{d(X+x_0)}_{dX}$$

$\Rightarrow \text{prob}(X|I) \approx \text{const. in allowed range}$

\Rightarrow complete ignorance : uniform pdf

QUANTITIES ASSOCIATED WITH A SIZE : SCALE PARAMETERS

$$\text{prob}(L|I) dL \approx \text{prob}(\beta L|I) d(\beta L) = \beta dL$$

$L \rightarrow \beta L$ should not result in change

$$\Rightarrow \text{prob}(L|I) \propto \frac{1}{L}$$

(Jeffrey's prior)

equivalent to uniform pdf for logarithm of
 $\text{prob}(\log L|I) = \text{constant}$

5.2. TESTABLE INFORMATION:

THE PRINCIPLE OF MAXIMUM ENTROPY

SO FAR: RELIED ON TRANSFORMATION GROUPS

SUPPOSE A DICE IS ROLLED VERY LARGE NUMBER OF TIMES

WE ARE TOLD: AVERAGE: 4.5

WHAT PROB SHOULD WE ASSIGN:

$$\sum_{i=1}^6 i \cdot \text{prob}(x_i | I) = 4.5$$

uniform pdf would result in 3.5 ⚡

WHICH pdf is the best ?

EXAMPLE FOR TESTABLE INFORMATION:

ACCEPT OR REJECT ANY PROPOSED PDF

JAYNES: PRINCIPLE OF MAXIMUM ENTROPY (Max Ent)
1957:

CHOOSE PDF WHICH HAS THE MOST ENTROPY

MAXIMIZE: $S = - \sum_{i=1}^6 p_i \ln p_i$

with $p_i = \text{prob}(x_i | I)$

subject to: $\sum_{i=1}^6 p_i = 1$ and $\sum_{i=1}^6 i p_i = 4.5$

constrained optimization: Lagrange multipliers

Example 1:

"normal" dice $S = - \sum_{i=1}^6 p_i \ln p_i$
 $\sum_{i=1}^6 p_i = 1$

Minimize $F = S + \lambda \left(\sum_{i=1}^6 p_i - 1 \right)$

$$\frac{\partial F}{\partial p_j} = -\ln p_j - 1 + \lambda \stackrel{!}{=} 0 \quad (1)$$

$$\frac{\partial F}{\partial \lambda} = \sum_{i=1}^6 p_i - 1 \stackrel{!}{=} 0 \quad (2)$$

From (1): $p_j = e^{\lambda-1} \quad (3)$

in (2): $\sum_{i=1}^6 e^{\lambda-1} - 1 = 0$

$$6 e^{\lambda-1} = 1$$

$$e^{\lambda-1} = \frac{1}{6}$$

$$\lambda - 1 = -\ln 6$$

$$\lambda = 1 - \ln 6$$

in (3): $p_j = e^{1 - \ln 6 - 1} = e^{-\ln 6} = \frac{1}{6} \quad \checkmark$

Example 2:

dice with $\sum_{i=1}^6 i p_i = 4.5$

$$\Rightarrow F = S + \lambda_1 \left(\sum_{i=1}^6 p_i - 1 \right) + \lambda_2 \left(\sum_{i=1}^6 i p_i - 4.5 \right)$$

$$\frac{\partial F}{\partial p_j} = -\ln p_j - 1 + \lambda_1 + j \lambda_2 \stackrel{!}{=} 0 \quad (1)$$

$$\frac{\partial F}{\partial \lambda_1} = \sum_{i=1}^6 p_i - 1 \stackrel{!}{=} 0 \quad (2)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_2} = \sum_{i=1}^6 i p_i - 4.5 = 0 \quad (2)$$

From (1): $p_j = \exp\{\lambda_1 + j\lambda_2 - 1\}$

in (2): $\sum_{i=1}^6 e^{\lambda_1 - 1} e^{i\lambda_2} = 1$

$$\sum_{i=1}^6 e^{i\lambda_2} = e^{1 - \lambda_1}$$

$$\ln \left[\sum_{i=1}^6 e^{i\lambda_2} \right] = 1 - \lambda_1$$

$$\Rightarrow \lambda_1 = 1 - \ln \left[\sum_{i=1}^6 e^{i\lambda_2} \right]$$

$$\Rightarrow p_i = \exp \left[-\ln \sum_{k=1}^6 e^{k\lambda_2} + i\lambda_2 \right]$$

$$= \frac{e^{i\lambda_2}}{\sum_{k=1}^6 e^{k\lambda_2}}$$

in (3):

$$\frac{\sum_{i=1}^6 i e^{i\lambda_2}}{\sum_{k=1}^6 e^{k\lambda_2}} = 4.5$$

$$\Leftrightarrow \sum_{i=1}^6 (i - 4.5) e^{i\lambda_2} = 0$$

Solve for $\lambda_2 \Rightarrow \lambda_2 = 0.37$

Lecture 6a. R

PLAUSIBILITY FOR MAXENT

(15)

(SEE MORE: E.T. JAYNES: PROBABILITY THEORY
THE LOGIC OF SCIENCE)

KANGAROO PROBLEM:

INFORMATION: $\frac{1}{3}$ of all kangaroos have blue eyes
 $\frac{1}{4}$ of kangaroos are left handed

QUESTION: WHAT PROPORTION OF KANGAROOS ARE
BOTH BLUE-EYED AND LEFT-HANDED?

FOR EACH KANGAROO 4 OPTIONS:

- 1) BLUE-EYED AND LEFT HANDED : P_1
- 2) BLUE-EYED AND RIGHT HANDED : P_2
- 3) NOT BLUE-EYED AND LEFT-HANDED : P_3
- 4) NOT BLUE-EYED AND RIGHT-HANDED : P_4

truth table

		left-handed	
		TRUE	FALSE
blue-eyed	FALSE	P_1	P_2
	TRUE	P_3	P_4

$0 \leq x \leq \frac{1}{4}$	$\frac{1}{3} - x$
$\frac{1}{4} - x$	$\frac{5}{12} + x$

Normalization: $P_1 + P_2 + P_3 + P_4 = 1$

CONDITIONS: $P_1 + P_2 = \frac{1}{3}$

$P_1 + P_3 = \frac{1}{4}$

3 eqn. for 4 variables
 \Rightarrow one unknown: x
(say P_1)

\Rightarrow solutions: $0 \leq x \leq \frac{1}{4}$

Which one is best?

From independence $x = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$

IS THERE A FUNCTION OF $\{p_i\}$ which when maximized to constraints yields this solution?

ONLY FUNCTION: $S = - \sum p_i \ln p_i$

others give correlations between handedness and eye-color (e.g. $-\sum p_i^2$, $\sum \ln p_i$, $\sum \sqrt{p_i}$)

$$= \left[x \cdot \ln x + \left(\frac{1}{3}-x\right) \ln \left(\frac{1}{3}-x\right) + \left(\frac{1}{4}-x\right) \ln \left(\frac{1}{4}-x\right) + \left(\frac{5}{12}-x\right) \ln \left(\frac{5}{12}-x\right) \right]$$

$$\frac{\partial S}{\partial x} = - \left[\ln x + 1 - \ln \left(\frac{1}{3}-x\right) - 1 - \ln \left(\frac{1}{4}-x\right) - 1 + \ln \left(\frac{5}{12}-x\right) + 1 \right]$$

$$= - \left[\ln x - \ln \left(\frac{1}{3}-x\right) - \ln \left(\frac{1}{4}-x\right) + \ln \left(\frac{5}{12}-x\right) \right]$$

$$= - \left[\ln \frac{x \left(\frac{5}{12}-x\right)}{\left(\frac{1}{3}-x\right) \left(\frac{1}{4}-x\right)} \right] = 0$$

$$x \left(\frac{5}{12}-x\right) = \left(\frac{1}{3}-x\right) \left(\frac{1}{4}-x\right)$$

$$\frac{5}{12}x - x^2 = \frac{1}{12} - \frac{1}{3}x + \frac{1}{4}x - x^2$$

$$\frac{5}{12}x + \frac{1}{3}x + \frac{1}{4}x = \frac{1}{12}$$

$$8x + 6x + 3x = 1 \Rightarrow x = \frac{1}{12} \quad \checkmark$$

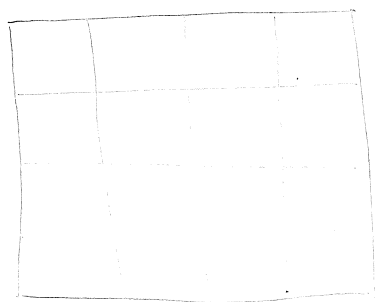
5.2.1 THE MONKEY ARGUMENT

SUPPOSE: M distinct possibilities $\{X_i\}$

assign truth values : $\text{prob}(X_i | I) = p_i$

How do we do this fairly?

Each possibility represented by a box of same size:



Assume monkeys throw pennies
at random in boxes

for a large number of coins
fraction in boxes \cong probability p_i

(if these p_i 's violate constraints
 $I \Rightarrow$ rejected)

Assume this experiment is repeated many times \Rightarrow
peaks around some distributions p_i

Quantify:

n_i : number of coins in i -th box

total no. of coins : $N = \sum_{i=1}^M n_i$

assume: $N \gg M$

$$p_i = \frac{n_i}{N}$$

Since every penny can land in any box: M^N different
ways

many of these same distribution $\{n_i\}$

$$F(\{p_i\}) = \frac{\text{number of ways of obtaining } \{n_i\}}{M^N}$$

(2)

Start with box 1:

In how many ways can n_1 coins be chosen from a total of N ?

ANSWER: $N C_{n_1}$

THEN BOX 2: In how many ways can n_2 be chosen from $N - n_1$: $N - n_1 C_{n_2}$

$$\Rightarrow \text{NUMERATOR: } N C_{n_1} \times N - n_1 C_{n_2} \times N - n_1 - n_2 C_{n_3} \times \dots \times n_m C_{n_m}$$

$$= \frac{N!}{n_1! n_2! n_3! \dots n_m!}$$

$$\Rightarrow \ln F = -N \ln M + \ln N! - \sum_{i=1}^M \ln n_i!$$

Stirling approximation: $\ln n! \approx n \ln n - n$

$$\Rightarrow \ln F \approx -N \ln M + N \ln N - \sum_{i=1}^M n_i \ln n_i$$

(used $\sum n_i = N$) together with $\sum p_i = 1$

$$\Rightarrow \ln F = -N \ln M - N \sum_{i=1}^M p_i \ln p_i$$

$$= - \sum_{i=1}^M n_i \ln n_i + \sum_{i=1}^M n_i \ln N = - \sum_{i=1}^M n_i (\ln n_i - \ln N)$$

$$= - \sum_{i=1}^M n_i \ln \frac{n_i}{N} = - \sum_{i=1}^M n_i \ln p_i$$

$$= - N \sum_{i=1}^M \frac{n_i}{N} \ln p_i = - N \sum_{i=1}^M p_i \ln p_i$$

The monkey experiment: Maximize $\ln F$

N and M are constant \Rightarrow

$$\Rightarrow \text{maximize: } S = - \sum_{i=1}^M p_i \ln p_i$$

Jaynes:
Prob.
the Logic
of Science!

5.2.2. THE LEBESGUE MEASURE

3

SO FAR: ALL BOXES SAME SIZE

SUPPOSE PROBLEM OF DIE, FOR A STRANGE REASON,
IS POSED AS

$X_i \equiv$ the face on top has $\begin{cases} i \text{ dots for } i=1,2 \\ 3,4,5 \text{ or } 6 \text{ dots for } i=3 \end{cases}$

\Rightarrow inclined to make box for X_3 4 times larger

\Rightarrow include this in previous analysis

chance that monkey throws penny in i -th box
is m_i : condition: $\sum_{i=1}^M m_i = 1$

$$\Rightarrow F(\{p_i\}) = \frac{N!}{n_1! n_2! \dots n_M!} \times m_1^{n_1} m_2^{n_2} \dots m_M^{n_M}$$

(if all m_i are equal: $m_i = \frac{1}{M}$)

multinomial distribution (binomial: $M=2$)

with Stirling's approximation:

$$\ln F = \sum_{i=1}^M n_i \ln m_i - N \sum_{i=1}^M p_i \ln p_i$$

$$\Rightarrow \frac{1}{N} \ln F = - \sum_{i=1}^M p_i \ln \frac{p_i}{m_i} \equiv S$$

SHANNON - JAYNES ENTROPY

JAYNES (1963): REQUIRED FOR CONTINUOUS PARAMETERS

$$S = - \int p(x) \ln \frac{p(x)}{m(x)} dx$$

$m(x)$: Lebesgue measure

ensures entropy is invariant under exchange of variables

$$x \rightarrow y = f(x)$$

because both $p(x)$ and $m(x)$ transform in the same way!

Example: only normalization condition $\int p(x) dx = 1$

for discrete counterpart:

$$Q = - \sum_i p_i \ln \frac{p_i}{m_i} + \lambda (1 - \sum_i p_i)$$

-(1+2)

$$\frac{\partial Q}{\partial p_j} = -1 - \ln \frac{p_j}{m_j} - \lambda = 0 \Rightarrow p_j = m_j e^{-\lambda}$$

$$\frac{\partial Q}{\partial \lambda} = \sum_i p_i - 1 \stackrel{!}{=} 0$$

$$\sum_i m_i e^{-\lambda} = 1 \Rightarrow e^{-\lambda} \sum_i m_i = 1$$

$$e^{-\lambda} = 1 \Rightarrow \lambda = 0$$

$$\Rightarrow p_j = m_j$$

$$\Rightarrow p(x) = \text{prob}(x | \text{norm}) \propto m(x) \text{ ~~const~~}$$

$m(x)$ is multiple of $p \Rightarrow$ complete ignorance of x !

$1/M \Rightarrow$ uniform

5.3 Max Ent examples: some common pdfs

Variance and Gaussian distribution

μ and it's variance known

$$\langle (x - \mu)^2 \rangle = \int (x - \mu)^2 \text{prob}(x|I) dx = \sigma^2$$

discrete case:

$$Q = - \sum_i p_i \ln \frac{p_i}{m_i} + \lambda_0 (1 - \sum_i p_i) + \lambda_1 (\sigma^2 - \sum_i (x_i - \mu)^2 p_i)$$

LAGRANGE
MULTIPLIERS

~~$$Q = - \sum_i p_i \ln \frac{p_i}{m_i}$$~~

$$\frac{\partial Q}{\partial p_j} = 0 \Rightarrow p_j = m_j e^{-(1+\lambda_0)} e^{-\lambda_1 (x_j - \mu)^2}$$

for uniform measure

$$\text{prob}(x|I) \propto \exp[-\lambda_1 (x - \mu)^2]$$

~~$$\frac{\partial Q}{\partial p_j} = - \ln \frac{p_j}{m_j} - p_j \left[\frac{1}{p_j} \right]$$~~

~~$$- \sum_i p_i (\ln p_i - \ln m_i)$$~~

~~$$\frac{\partial Q}{\partial p_j} = - \ln \frac{p_j}{m_j} - 1 = \lambda_0 - \lambda_1 (x_j - \mu)^2 = 0$$~~

~~$$\Rightarrow p_j = m_j e^{\lambda_0 - 1} e^{-\lambda_1 (x_j - \mu)^2}$$~~

assume m_j uniform:

$$\sum_i p_i = 1$$

$$\langle (x^2 - \bar{x})^2 \rangle = \sigma^2$$

$$\left. \begin{array}{l} \sum_i p_i = 1 \\ \langle (x^2 - \bar{x})^2 \rangle = \sigma^2 \end{array} \right\} \Rightarrow p(x|\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

EXTENSION TO MULTIPLE VARIABLES:

$$S = - \iint \dots \int p(\vec{x}) \ln \frac{p(\vec{x})}{m(\vec{x})} d^N x$$

$$\langle (x_k - \mu_k)^2 \rangle = \iint \dots \int (x_k - \mu_k)^2 p(\vec{x}) d^N x = \sigma_k^2$$

Uniform measure:

$$\text{prob}(\{x_k\} | \{\mu_k, \sigma_k\}) = \prod_{k=1}^N \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left[-\frac{(x_k - \mu_k)^2}{2\sigma_k^2}\right]$$

Max Ent and the binomial distribution

number of successes in M trials: $\langle N \rangle = \mu$

What is the probability of successful outcomes:

$$\text{prob}(N | M, \mu)$$

$$\text{TESTABLE INFORMATION: } \langle N \rangle = \sum_{N=0}^M N \text{prob}(N | M, \mu) = \mu \quad (*)$$

(+ NORMALIZATION)

$$\text{if mean is assigned: } \text{prob}(N | M, \mu) \propto m(N) e^{-\lambda N}$$

(HOMEWORK!)

λ : Lagrange multiplier

$m(N)$: measure

$m(N)$: measure which reflects gross ignorance!

equal probability to 2^M possible outcomes

$$\Rightarrow m(N) = \frac{M!}{N! (M-N)!}$$

⇒ NORMALIZATION:

$$\sum_{N=0}^M A m(N) e^{-\lambda N} = A (e^{-\lambda} + 1)^M \stackrel{!}{=} 1$$

(BINOMIAL EAM)

A: normalization

$$\Rightarrow A = \frac{1}{(e^{-\lambda} + 1)^M}$$

$$\Rightarrow \text{prob}(N|M, \mu) = \frac{1}{(e^{-\lambda} + 1)^M} \frac{M!}{N! (M-N)!} e^{-\lambda N}$$

use $\sum_{N=0}^M m(N) e^{-\lambda N} = (e^{-\lambda} + 1)^M$

derivative w.r.t λ : (both sides)

$$-\sum_{N=0}^M N m(N) e^{-\lambda N} = M (e^{-\lambda} + 1)^{M-1} (-1) e^{-\lambda}$$

$$\Rightarrow \text{in (*)} \quad \sum_{N=0}^M N \frac{m(N) e^{-\lambda N}}{(1+e^{-\lambda})^M} = \frac{M (1+e^{-\lambda})^{M-1} e^{-\lambda}}{(1+e^{-\lambda})^M} \stackrel{!}{=} \mu$$

$$\Leftrightarrow M (1+e^{-\lambda})^{-1} e^{-\lambda} = \mu$$

$$\Leftrightarrow \frac{M}{(1+e^{-\lambda}) e^{\lambda}} = \frac{M}{(e^{\lambda} + 1)} \stackrel{!}{=} \mu$$

$$\Rightarrow \ln\left(\frac{M}{\mu} - 1\right) = \lambda$$

$$\Rightarrow \text{prob}(N|M, \mu) = \frac{1}{\left(\frac{1}{\mu/\mu - 1} + 1\right)^M} \frac{M!}{N!} \frac{1}{\left(\frac{M}{\mu} - 1\right)^N} \frac{1}{(M-N)!} \quad (8)$$

$$= \frac{1}{\left(\frac{\mu/\mu}{\mu/\mu - 1}\right)^M} \frac{M!}{N!} \left(\frac{M}{\mu} - 1\right)^{-N} \frac{1}{(M-N)!}$$

$$= \frac{M!}{N!(M-N)!} \left(\frac{\mu}{M}\right)^N \left(1 - \frac{\mu}{M}\right)^{M-N}$$

COUNTING AND POISSON STATISTICS

IN GENERAL: COUNTING DISCRETE EVENTS IN A FINITE INTERVAL

WHAT IS PROBABILITY TO OBSERVE N EVENTS

$$\langle N \rangle = \mu$$

LARGE NUMBER OF MICROSCOPIC SUB-INTERVALS WHERE A SINGLE EVENT CAN OCCUR.

LIKE $M \rightarrow \infty$ BEFORE !

SHOWN LATER!
FOR NOW MAX ENT

$$m(N) = \frac{M!}{N!(M-N)!} \xrightarrow{M \rightarrow \infty} \frac{M^N}{N!}$$

see before
 \Rightarrow

$$\text{prob}(N|\mu) = A \frac{(M e^{-\lambda})^N}{N!}$$

TAYLOR
OF EXP

NORMALIZATION

$$\Rightarrow \sum_{N=0}^{\infty} \text{prob}(N|\mu) = A e^{M e^{-\lambda}} \stackrel{!}{=} 1$$

$$\Rightarrow \text{prob}(N|\mu) = \frac{1}{e^{\mu e^{-\lambda}}} \frac{(\mu e^{-\lambda})^N}{N!}$$

9

$$\langle N \rangle = \mu \Rightarrow \mu e^{-\lambda} = \mu$$

$$\Rightarrow \text{prob}(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$

5.4. APPROXIMATIONS : INTER CONNECTIONS AND

SIMPLIFICATIONS

BINOMIAL DISTRIBUTION IN THE LARGE M LIMIT

$$\text{prob}(N|M, \mu) \approx \frac{M^N}{N!}$$

$$\frac{M!}{(M-N)!} \approx \overbrace{M \times M \times \dots \times M}^{N\text{-mal}} \approx M^N$$

$$\text{Further: } \left(1 - \frac{\mu}{M}\right)^{M-N} \approx \left(1 - \frac{\mu}{M}\right)^M$$

$$\Rightarrow \text{prob}(N|M, \mu) \approx \frac{M^N}{N!} \left(\frac{\mu}{M}\right)^N \underbrace{\left(1 - \frac{\mu}{M}\right)^M}_\rightarrow e^{-\mu} \text{ for } M \rightarrow \infty$$

binomial pdf:

$$\text{prob}(r|N, p) = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r} \approx \underbrace{\frac{(Np)^r e^{-Np}}{r!}}_{\text{Poisson}}$$

for small p and large N

EXAMPLE: ROTTEN APPLES

NUMBER: $N=100$ $R: \text{dbinom}(r, 100, 0.02)$
 $p=0.02$

POISSON: $\mu = Np = 2$ $r = 0, \dots, 7$ $r=0: 0.135$ bin
 $R: \text{dpois}(r, 2)$ 0.133 coin

FURTHER APPROXIMATION

LOGARITHM OF POISSON DISTRIBUTION:

$$L = \ln(\text{prob}(N|\mu)) = N \ln \mu - \mu - \ln N!$$

Most probable values of N close to μ for $\mu \gg 1$ Stirling's formula:

$$\ln N! \approx N \ln N - N + \frac{1}{2} \ln(2\pi N)$$

negligible for $N \rightarrow \infty$
 keep for the moment!

$$\Rightarrow L \approx N \ln \mu - \mu - N \ln N + N - \frac{1}{2} \ln(2\pi N)$$

$$= N - \mu - N \ln \frac{N}{\mu} - \frac{1}{2} \ln(2\pi N)$$

 N in neighborhood of μ : $N = \mu + \epsilon$ with $|\epsilon| \ll \mu$:

$$L \approx -\frac{1}{2} \ln(2\pi \mu) + \epsilon - \left(\mu + \epsilon + \frac{1}{2}\right) \ln\left(1 + \frac{\epsilon}{\mu}\right)$$

why?

$$\epsilon - \left(\mu + \epsilon\right) \ln\left(1 + \frac{\epsilon}{\mu}\right) - \frac{1}{2} \ln\left[2\pi\left(\mu + \epsilon\right)\right]$$

$$-\frac{1}{2} \ln \left(2\pi\mu \left(1 + \frac{\epsilon}{\mu} \right) \right)$$

$$= -\frac{1}{2} \left[\ln 2\pi\mu + \ln \left(1 + \frac{\epsilon}{\mu} \right) \right] \quad \text{ok}$$

TAYLOR EXPAND $\ln \left(1 + \frac{\epsilon}{\mu} \right)$

$$\Rightarrow L = -\frac{1}{2} \ln(2\pi\mu) - \frac{\epsilon^2}{2\mu} - \frac{\epsilon}{2\mu} + \dots$$

$$= -\frac{1}{2} \ln(2\pi\mu) - \frac{1}{2\mu} (\epsilon^2 - \epsilon)$$

$$= -\frac{1}{2} \ln(2\pi\mu) - \frac{1}{2\mu} \left[\cancel{(N+\mu)} (N-\mu)^2 - N + \mu \right]$$

$$= -\frac{1}{2} \ln(2\pi\mu) - \frac{1}{2\mu} \left[N^2 - 2\mu N + \mu^2 - N + \mu \right]$$

$$= -\frac{1}{2} \ln(2\pi\mu) - \frac{1}{2\mu} \left(\epsilon^2 + \epsilon + \frac{1}{4} - \frac{1}{4} \right)$$

$$= -\frac{1}{2} \ln(2\pi\mu) - \frac{1}{2\mu} \left(\epsilon + \frac{1}{2} \right) + \frac{1}{8\mu}$$

negligible

$$\Rightarrow \text{prob}(N|\mu) = \frac{\mu^N e^{-\mu}}{N!} \approx \frac{1}{\sqrt{2\pi\mu}} \exp \left[-\frac{(N-\mu+\frac{1}{2})^2}{2\mu} \right]$$

$$\text{also } \frac{\mu^N e^{-\mu}}{N!} \approx \frac{1}{\sqrt{2\pi N}} \exp \left[-\frac{(\mu-N)^2}{2N} \right]$$

(also corresponding limit of Binom of course!)

$$\mu = Np; \quad \sigma^2 = Np(1-p)$$

6 NON-PARAMETRIC ESTIMATION

- IF NO FUNCTIONAL MODEL EXISTS

6.1. INTRODUCTION: FREE-FORM SOLUTIONS

GIVEN SOME DATA $\{D_k\}$ and BACKGROUND INFO I
WHAT CAN WE SAY ABOUT OBJECT OF INTEREST $f(x)$?

$$\text{prob}[f(x) | \{D_k\}, I]$$

IF WE KNOW A LOT ABOUT $f(x)$ WE CAN SPECIFY
IT BY A FEW SPECIFIC PARAMETER

FOR EXAMPLE: $f(x)$ one-dimensional (spectrum

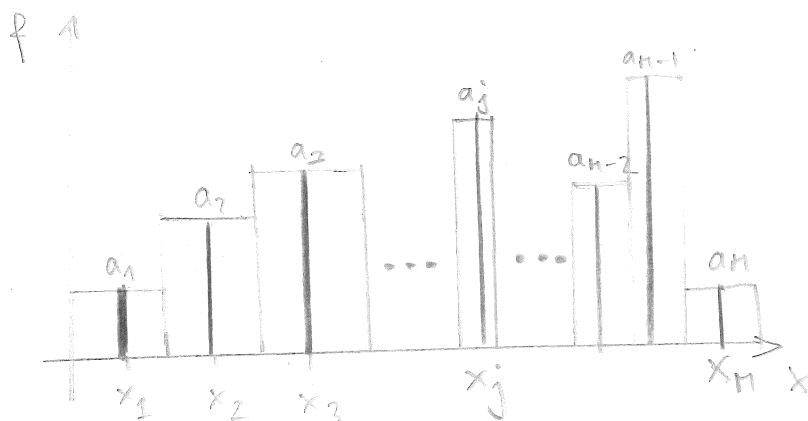
$$f(x) = a_1 \delta(x-x_1) + a_2 \delta(x-x_2)$$

$\delta(x-x^*)$: sharp spike at $x=x^*$

$$\Rightarrow \text{prob}(a_1, a_2, x_1, x_2 | \{D_k\}, I)$$

(SIMPLE FOUR PARAMETER MODEL)

~~LET~~ NEXT EXAMPLE: FEW SHARP PEAS, EXACT
NUMBER UNKNOWN



$$f(x) = \sum_{j=1}^M a_j \delta(x-x_j)$$

HOW MANY PEAKS?

CALCULATE: $\text{prob}(M | \{D_n\}, I) \leftarrow \text{USE MODEL SELECTION}$

MORE DIFFICULT: SO LITTLE KNOWLEDGE THAT FUNCTIONAL FORM IS NOT CHARACTERIZED

\Rightarrow ALLOW FOR A LOT OF FLEXIBILITY

\Rightarrow LARGE NUMBER OF PARAMETERS

\hookrightarrow PARTICULAR PROBLEM: CHOICE OF PRIOR

\hookrightarrow ANY WEAK CONSTRAINT IS USEFUL

6.1.1 SINGULAR VALUE DECOMPOSITION

EXAMPLE: ONE-DIMENSIONAL SPECTRUM

DIVIDE X-AXIS IN TINY PIECES

AVERAGE VALUE OF f IN EACH BIN

(SEE PREVIOUS FIGURE)

$$f(x) = \sum_{j=1}^M a_j \delta(x - x_j)$$

M: LARGE

$$x_j = x_{\min} + \left[\frac{x_{\max} - x_{\min}}{M} \right] \left(j - \frac{1}{2} \right)$$

$\Rightarrow \text{prob}(\{a_j\} | \{D_n\}, I)$

DIVISION OF FUNCTION IN DISCRETE "PIXELS"

(EXTENSION TO MORE DIMENSIONS)

LIKELIHOOD FUNCTION : $\text{prob}(\{D_k\}|\{a_j\}, I)$

FOR WELL DEFINED PEAK UNIFORM PRIOR SUFFICIENT

$$\ln[\text{prob}(\{D_k\}|\{a_j\}, I)] = \text{const.} - \frac{1}{2} \sum_{k=1}^N \frac{(F_k - D_k)^2}{\sigma_k^2}$$

F_k : prediction based on given $\{a_j\}$

ASSUME: ~~DATA~~ F_k RELATED LINEARLY TO " $f(x)$ "

$$F_k = \sum_{j=1}^M T_{kj} a_j + C_k$$

ELEMENTS OF MATRIX T AND \vec{C} INDEPENDENT OF a_j !

EXAMPLE: T_{kj} equal to the sine of $\frac{2\pi jk}{M}$ FOR INTERFEROMETRIC MEASUREMENTS
 \vec{C} : SLOWLY VARYING BACKGROUND

LIKELIHOOD FUNCTION IS ELLIPTICAL

M-DIMENSIONAL ELLIPSOID :

DIRECTIONS AND WIDTH OF PRINCIPAL AXES ARE GIVEN BY EIGENVECTORS $\{\vec{e}_e\}$ AND EIGENVALUES $\{\lambda_e\}$ OF $M \times M$ MATRIX OF SECOND DERIVATIVES OF $\ln[\text{prob}(\{D_k\}|\{a_j\}, I)]$ WHICH IS $\vec{\nabla} \vec{\nabla} \chi^2$

SOLVE $\vec{\nabla} \vec{\nabla} \chi^2 \vec{e}_e = \lambda_e \vec{e}_e$

FOR $L = 1, 2, \dots, M$

$$[\vec{\nabla} \vec{\nabla} x^2]_{ij} = \frac{\partial^2 x^2}{\partial a_i \partial a_j} = 2 \sum_{k=1}^N \frac{T_{ki} T_{kj}}{a_k^2}$$

WIDTH OF PROBABILITY ELLIPSOID INVERSELY
PROPORTIONAL TO EIGENVALUES

IF EIGEN VALUE IS VERY SMALL OR EVEN ZERO \Rightarrow
 PARAMETER NOT CONSTRAINT \Rightarrow FLAT DIRECTION OF
 $\text{prob}(\{D_k\}|\{a_{ij}\}, I)$

ALWAYS WHEN $m > n$!

* "EASY" CHECK: \det of $\vec{\nabla} \vec{\nabla} x^2$ is (almost) zero

↳ SINGULAR MATRIX

ANALYSIS OF EIGEN PROPERTIES: SINGULAR VALUE DECOMPOSITION!

large λ_2 : $X+Y$ well constrained
small λ_1 : $X-Y$ NOT

~~THE~~ EIGEN VECTOR IS: (α, β)

(5)

REFERS TO QUANTITY: $\alpha X + \beta Y$ (FOR 2-d problem)

IN GENERAL:

M ELEMENTS OF \vec{e}_e refer to different linear combinations of $\{a_e\}$

QUANTITIES ARE UNCORRELATED, SINCE \vec{e}_e are \perp TO EACH OTHER, \Rightarrow ESTIMATED INDEPENDENTLY!

\Rightarrow EIGEN^{VECTORS}~~VALUES~~ WITH LARGE λ REPRESENT ASPECTS OF $f(x)$ WHICH CAN BE "WELL" CONSTRAINED

EXAMPLE

GAUSSIAN: $T_{kj} = \frac{1}{w\sqrt{2\pi}} \exp\left[-\frac{(x_k - x_j)^2}{2w^2}\right]$

x_k : k-th data point

x_j : j-th pixel (bin)

Choose same grid and data: $M = N = 64$

WITH ALL ~~64~~

Lecture 8a.R

6.1.2. A parametric free-form solution?

⑥

A FREE FORM SOLUTION COULD CONTAIN LARGE NUMBER OF FREE PARAMETERS.

USE FORM WHICH MIGHT CONTAIN FEWER FREE VARIABLES:

APPROXIMATIONS FOR EXAMPLE BY TAYLOR SERIES

$$f(x) = c_1 + c_2(x-x_0) + c_3(x-x_0)^2 + c_4(x-x_0)^3 + \dots$$

IF NECESSARY ONE COULD CARRY OUT MODEL SELECTION TO DETERMINE OPTIMAL EXPANSION ORDER.

IN GENERAL EXPANSION INTO BASIS FUNCTIONS $\eta_L(x)$

$$f(x) = \sum_{L=1}^{M_b} c_L \eta_L(x)$$

TAYLOR: $\eta_L(x) = (x-x_0)^{L-1}$

SYMMETRY OF THE PROBLEM:

Chebyshev polynomials
spherical harmonics
sinusoids
wavelets
⋮

BEFORE: $\eta_L(x) = \delta(x-x_L)$

AIM: M_b SHOULD BE SMALL

EIGENVECTORS OF $\vec{\nabla} \vec{\nabla} \chi^2$ NATURAL CHOICE

$$\eta_L(x) = \vec{e}_L(x)$$

WITH OMISSION OF EKS WITH SMALL EIGENVALUES!

$$f(x_j) = a_j = \sum_{e=1}^{M_b} c_e \vec{e}_e(x_j)$$

FOR $j = 1, \dots, M$

THE SUM IS ONLY OVER THOSE e 's that have LARGE λ_e 's : $M_b \ll M$ (TYPICALLY)

$$c_e = \frac{2}{\lambda_e} \sum_{k=1}^N \frac{D_k}{6_k^2} \sum_{j=1}^M T_{kj} e_e(x_j) \pm \sqrt{\frac{2}{\lambda_e}}$$

(assuming orthonormalized EIGENVECTORS)

VERY IMPORTANT: INCLUDE EVERY PRIOR INFORMATION $\text{prob}[f(x) | \vec{T}]$

6.2. MAXENT: IMAGES, MONKEYS AND A NON-UNIFORM PRIOR

OBJECT OF INTEREST: POSITIVE AND ADDITIVE
 $\Rightarrow f(x) \geq 0$
 INTEGRAL (OR SUM) HAS
 "PHYSICAL" MEANING

WHAT PRIOR TO CHOOSE?

LAST CHAPTER,

MONKEY ARGUMENT: 2-d grid, blobs (luminiscent)
 \Rightarrow result: IMAGE

Eqn 5.28 :

$$S = - \sum_{i=1}^M p_i \ln \left(\frac{p_i}{m_i} \right) -$$

Suggests

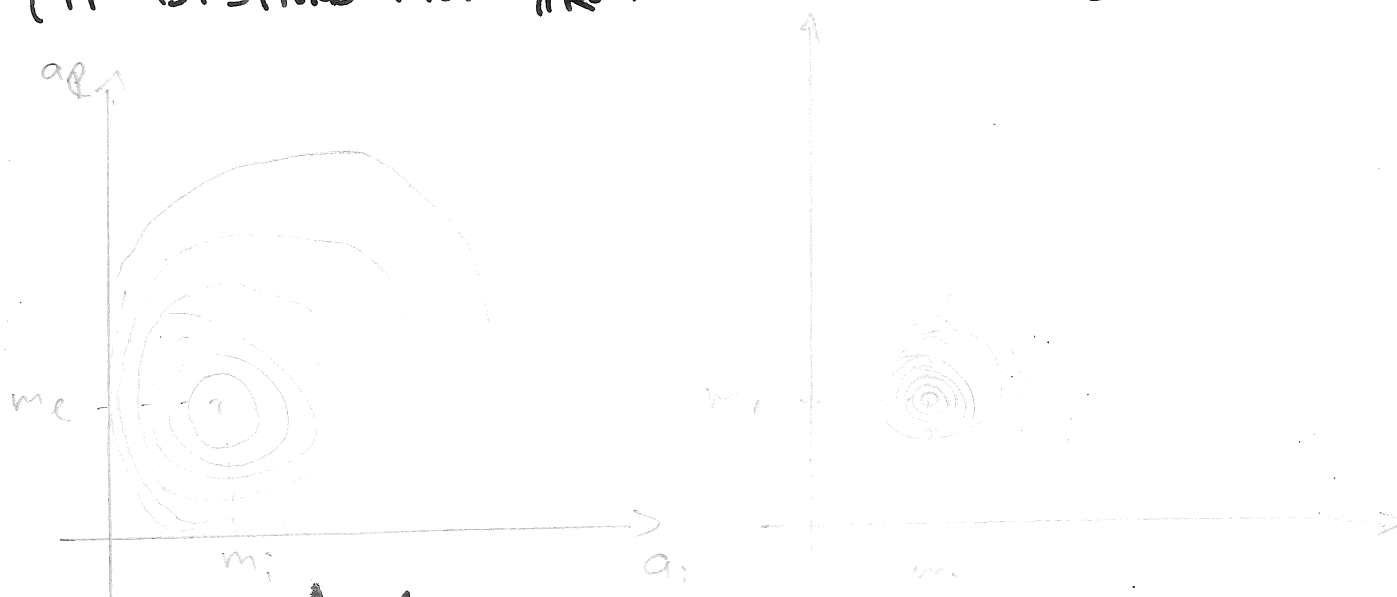
$$\text{prob}(\{a_j\} | \{m_j\}, \alpha, I) \propto \exp(\alpha S)$$

α : constant

$\{a_j\}$ relative to Lebesgue measure m_j

$$S = \sum_{j=1}^M \left(a_j - m_j - a_j \left[\ln \left[\frac{a_j}{m_j} \right] \right] \right)$$

(IF DISTRIBUTIONS ARE NOT NORMALIZED)



$\alpha = 1$
CONTOURS OF CONSTANT
 $\exp(\alpha S)$

$\alpha = 10$

($\alpha \sim$ number of balls)

Maximum at $\{a_j\} = \{m_j\}$

TOGETHER WITH GAUSSIAN LIKELIHOOD FROM BEFORE

⑨

$$\text{pro}(\{a_{ij}\} | \{D_k\}, \{m_{ij}\}, \alpha, I) \propto \exp\left[\alpha S - \frac{\chi^2}{2}\right]$$

HOWEVER: RESULT DEPENDS ON DEGREE OF
PIXELIZATION \ddots

6.2.1 REGULARIZATION

MOST PROBABLE ESTIMATE FOR $\{a_{ij}\}$

MAXIMUM OF $\alpha S - \frac{\chi^2}{2}$

LIKE CONSTRAINED MINIMIZATION WITH α LAGRANGE MULTIPLIER.

\Rightarrow S IS REGULARIZATION FUNCTION WHICH HELPS TO STABILIZE LEAST-SQUARE PROCEDURE FOR FREE-FORM SOLUTION!

$$\frac{\partial}{\partial a_{ij}} \left(\alpha S - \frac{\chi^2}{2} \right) = 0 \quad j = 1, 2, \dots, M$$

NUMERICALLY DEMANDING!

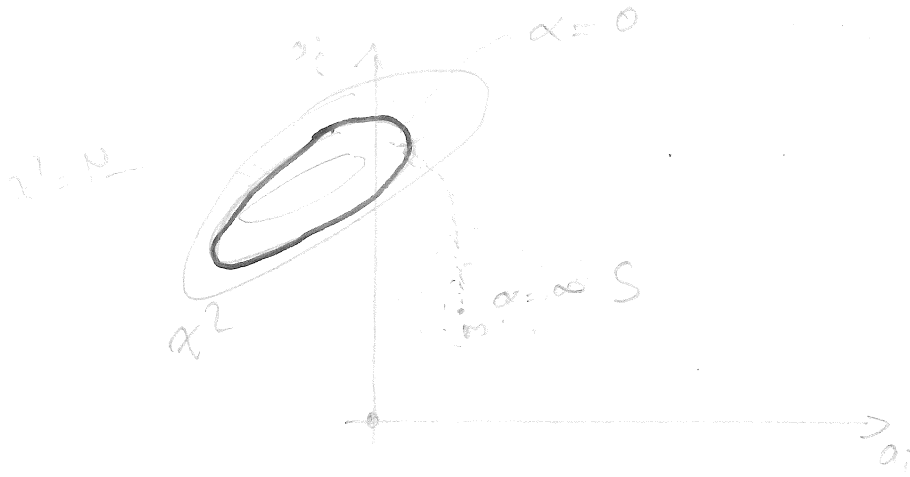
ITERATIVE SCHEME (Skilling & Bryan 1984):

START: $\{a_{ij}\} = \{m_{ij}\}$ (GLOBAL MAX WITHOUT DATA)

\hookrightarrow LOOK AT LOCAL GRADIENT OF S AND χ^2

\Rightarrow BEST SMALL CHANGE: $\{\Delta a_{ij}\}$

\Rightarrow REPEAT UNTIL: $2\alpha \vec{\nabla} S = \vec{\nabla} \chi^2$



MaxEnt trajectory

CONDITIONS: NO MULTIMODAL LIKELIHOOD

CHOOSE AS STEPLENGTH:

$$\Delta R^2 = \sum_{j=1}^M \frac{\Delta a_j^2}{a_j}$$

6.3. SMOOTHNESS: FUZZY PIXELS AND SPATIAL CORRELATIONS

INCLUDE IN PRIOR: OBJECT IS LOCALLY SMOOTH!

GO BACK TO BEGINNING: FREE FORM WAS

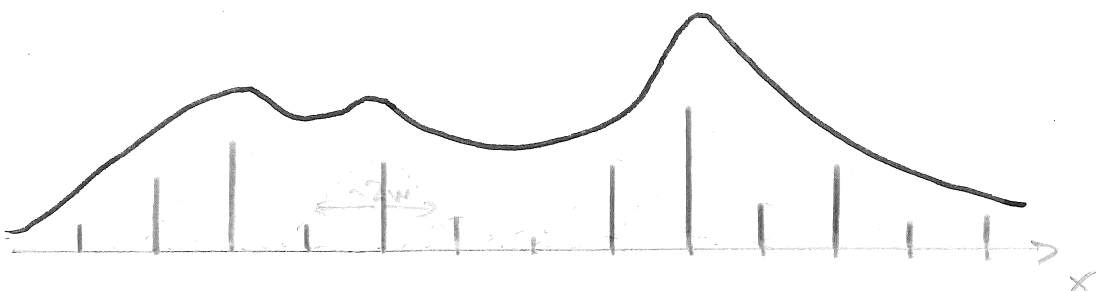
$$\eta_j(x) = \delta(x - x_j)$$

BETTER:

$$f(x) = \sum_{j=1}^M a_j \exp\left[-\frac{(x - x_j)^2}{2w^2}\right]$$

"SPATIAL" EXTEND \Rightarrow SMOOTHNESS

FUZZY PIXELS!



HOW DO WE CHOOSE w ?

DO WE NEED SPECIFIC BASISFUNCTION SETUP? NO

REGULARIZATION: LET w AS LARGE AS POSSIBLE
WHILE SATISFYING DATA CONSTRAINTS

~~6.4.1. INTERPOLATION~~

~~GIVEN ACTUAL VALUES OF $f(x)$ AT COARSE INTERVALS~~

~~PROBLEM~~

6.4.2. INFERENCE OR INVERSION?

$$\underset{\substack{\uparrow \\ \text{data}}}{\vec{F}} = \underset{\substack{\uparrow \\ \text{experiment}}}{T} \underset{\substack{\uparrow \\ \text{pixel}}}{\vec{a}} + \underset{\substack{\uparrow \\ \text{noise}}}{\vec{C}} \quad \text{(FOR LINEAR RELATION)}$$

INVERSE OPERATOR

$$\vec{a} = T^{-1} (\vec{F} - \vec{C})$$

T NEEDS TO BE SQUARE MATRIX

USUALLY NO. OF PIXELS M MUCH LARGER THAN DATA N .

EXAMPLE: SIMPLE CONVOLUTION PROBLEM

$$\text{DATA: } D(x_k) = \int f(x) R(x_k - x) dx + B(x_k) \pm G(x_k)$$

FT (convolution)

$$\hat{D}(w) \approx \hat{f}(w) \hat{R}(w) + \hat{B}(w)$$

$\hat{f}(w)$ slowly
modulated
background
 $\hat{B}(w)$ noise of
measurement

$$\hat{f}(w) = \int f(x) \exp(i2\pi x) dx$$

$$\Rightarrow f(x) \approx \int \left[\frac{\hat{D}(\omega) - \hat{B}(\omega)}{\hat{R}(\omega)} \right] \exp(-2i\pi\omega x) d\omega$$

SHOW FIG 6.10 ← MAY BE PRODUCE LATER

$f(x)$ and D_n on 128 points

6.10 (d) NOT AS INFORMATIVE AS DATA !!!

RETAINING $\delta(x_n)$:

$$\int \left[\frac{\hat{D}(\omega) - \hat{B}(\omega)}{\hat{R}(\omega)} \right] \exp(-i2\pi\omega x) d\omega = f(x) \pm \epsilon(x)$$

with $\epsilon(x) \approx \int \left[\frac{\hat{\delta}(\omega)}{\hat{R}(\omega)} \right] \exp(-2i\pi\omega x) d\omega$

DOMINATED BY SMALL FOURIER COMPONENTS OF R !
 SINCE $\hat{R}(\omega) \rightarrow 0$ for large $\omega \Rightarrow$ ARTIFACTS
 FROM HIGH FREQUENCY.

$$f(x) \approx \int \left[\frac{\hat{D}(\omega) - \hat{B}(\omega)}{\hat{R}(\omega)} \right] \hat{Q}(\omega) \exp(-i2\pi\omega x) d\omega$$

↑
SMOOTHING FUNCTION

$\hat{Q}(\omega) \approx 1$ BUT $\rightarrow 0$ AS LEAST AS FAST
 AS $\hat{R}(\omega)$

\Rightarrow INVERSE IS USUALLY NOT USEFUL!

7 EXPERIMENTAL DESIGN7.1. INTRODUCTION

SIMPLE EXAMPLE: COIN-FLIPPING

ONLY ONE WAY TO CHANGE EXPERIMENT:
NUMBER OF TOSSES

HOW MANY FLIPS N WILL BE REQUIRED
TO ESTIMATE H TO A GIVEN DEGREE OF
CONFIDENCE?

posterior pdf prob($H/N, R, I$)

(Section 2.2.) approximated error bar

$$\sigma = \sqrt{\frac{H_0(1-H_0)}{N}}$$

with: $H_0 = R/N$

\Rightarrow ANSWER DEPENDS ON BIAS OF COIN!

HOWEVER: FOR GIVEN N ERROR BAR WILL BE LARGEST
FOR $H_0 = 1/2$

$$\text{IF WE WANT } \sigma \leq \sigma_c \Rightarrow N \geq \frac{0.25}{\sigma_c^2}$$

$$\text{ERROR BAR} \sim \frac{1}{\sqrt{N}}$$

In general: posterior prob $(\{X_i\} | \{D_n\}, I)$

BAYES' :

$$\text{prob}(\{X_i\} | \{D_n\}, I) \propto \text{prob}(\{D_n\} | \{X_i\}, I) \times \text{prob}(\{X_i\}, I)$$

LIKELIHOOD

- IF SHARPLY PEAKED (MORE THAN PRIOR)
 \Rightarrow LEARNT A LOT FROM EXPERIMENT
- IF VERY BROAD (MORE THAN PRIOR)
 \Rightarrow LEARNT LITTLE

\Rightarrow MAKE LIKELIHOOD AS SHARPLY PEAKED AS POSSIBLE!

\hookrightarrow MAKE DATA MOST SENSITIVE TO QUANTITIES OF INTEREST!

7.2 EXAMPLE 7: OPTIMIZING RESOLUTION FUNCTIONS

③

SIMPLE CONVOLUTION PROBLEMS

RESOLUTION - OR POINT - SPREAD FUNCTION

(e.g. TELESCOPES)

ONE-DIMENSIONAL PROBLEM

resolution functions:

$$R(x) \propto \exp\left[-\frac{x^2}{2w^2}\right]$$

(in many situations, does not depend on position in x)

can also be highly asymmetric:

$$R(x) \propto e^{-x/\tau} \quad \text{for } x \geq 0$$

(zero otherwise)

IDEAL DATA:

$$F_k = T \int f(x) R(x_k - x) dx + B(x_k)$$

FOR $k=1, 2, \dots, N$ WITH $B(x)$ BACKGROUND SIGNAL

T : SCALING CONSTANT (could be \sim observing time)

INVESTIGATE SPREAD OF:

$$\text{prob}(\{D_k\} | f(x), T, R(x), B(x), I)$$

(as a function of experimental parameters)

example: counting

$$\text{prob}(\{D_k\} | \{F_k\}, I) = \prod_{k=1}^N \frac{F_k^{D_k} e^{-F_k}}{D_k!}$$

NEED TO SPECIFY $f(x)$

7.2.1 AN ISOLATED SHARP PEAK

$$f(x) = A \delta(x - \mu)$$

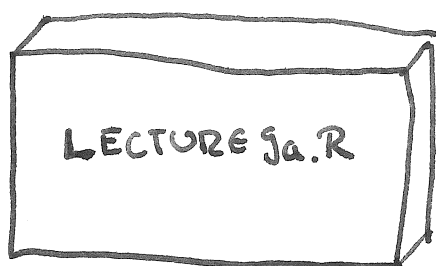
$$\Rightarrow F_k = A T R(x_k - \mu) + B(x_k)$$

NUMERICAL BRUTE FORCE METHOD

(TO ASSESS IMPACT OF EXPERIMENTAL PARAMETERS)

- i) CREATE MOCK DATA $\{D_k\}$
- ii) EVALUATE PROBABILITY FOR DIFF. VALUES A, μ
- iii) DISPLAY 2-dim pdf in A, μ space
- iv) EXPLORE HOW CONTOURS CHANGE WITH W or τ parameters (experimental)

GAUSSIAN RESOLUTION FUNCTION

TAKE $B(x) = B$ (constant)
 $\Delta = 1$
 $\mu = 0$

CONTOURS: 10, 30, 50, 70, 90%

- 1) $T = 20, w = 16, B = T/10$
- 2) $T = 200, w = 16, B = T/10 \rightarrow \sqrt{10} \sim 3$ narrower
- 3) $T = 20, w = 4, B = T/10$
 \rightarrow shrinks in μ but expands in A
 (R reduces no. of counts and blurring)
- 4) $T = 20, w = 16, B = T/2$
 (even worse for B marginalized)

THEORYSUPPOSE TRUE VALUES: $A = 1, \mu = 0$

ideal Data

$$\hat{D}_n = T \exp\left[-\frac{x_n^2}{2w^2}\right] + B$$

 $n = 1, 2, \dots, N$ add Gaussian noise: $D_n = \hat{D}_n + \epsilon_n$ with $\langle \epsilon_n \rangle = 0$ and $\langle \epsilon_n^2 \rangle = \sigma_n^2 = \hat{D}_n$ (Poisson)

$$\Rightarrow \ln [\text{prob}(\{D_n\} | \{F_n\}, I)] \approx \text{constant} - \chi^2/2$$

$$\Rightarrow \chi^2 = \sum_{k=1}^N \frac{(F_n - D_n)^2}{\sigma_n^2}$$

with

⑥

$$F_k = A T \exp\left[-\frac{(x_k - \mu)^2}{2w^2}\right] + B$$

MAXIMUM at

$$\frac{\partial \chi^2}{\partial A} = 0 \quad \frac{\partial \chi^2}{\partial \mu} = 0$$

$$A_0 \approx 1$$

$$\mu_0 \approx 0$$

Spread around optimal point

$$\frac{\partial^2 \chi^2}{\partial A^2} ; \frac{\partial^2 \chi^2}{\partial \mu^2} \quad \text{and} \quad \frac{\partial^2 \chi^2}{\partial A \partial \mu}$$

$$\frac{\partial^2 \chi^2}{\partial A^2} = \sum_{k=1}^N \frac{2T^2}{\sigma_k^2} \exp\left[-\frac{(x_k - \mu)^2}{2w^2}\right]$$

(evaluated at max $\mu_0 = 0$)

$$\Rightarrow \frac{\partial^2 \chi^2}{\partial A^2} \Big|_{A_0, \mu_0} \approx \sum_{k=1}^N \frac{2T^2}{\sigma_k^2} \exp\left[-\frac{x_k^2}{2w^2}\right]$$

Substituting $\sigma_k^2 = \hat{D}_k$ the sum can be done by an integral

$$\int_{-\infty}^{+\infty} \frac{\exp(-x^2/w^2)}{T \exp(-x^2/2w^2) + B} dx \approx \frac{w\sqrt{2\pi}}{T+B\sqrt{2}}$$

(asympt. correct in limits $B \rightarrow 0$, and $B \gg T$)
5% accuracy intermed. regime)

$$\Rightarrow \left. \frac{\partial^2 \chi^2}{\partial A^2} \right|_{A_0, \mu_0} \propto \frac{w T^2}{T + B\sqrt{2}}$$

Similar for other derivatives

$$\left. \frac{\partial^2 \chi^2}{\partial \mu^2} \right|_{A_0, \mu_0} \propto \frac{T^2}{w(T + B\sqrt{2})} \quad ; \quad \frac{\partial^2 \chi^2}{\partial A \partial \mu} \approx 0$$

THE COVARIANCE IS TWICE THE INVERSE OF $\vec{\nabla} \vec{\nabla} \chi^2$

OFF-DIAGONAL ELEMENT ZERO \Rightarrow NO CORRELATION BETWEEN A AND μ
(SEE ALSO FIGURES)

IN CASE OF SMALL BACKGROUND: $B \ll T$

$$\begin{pmatrix} \langle \delta \mu^2 \rangle & \langle \delta \mu \delta A \rangle \\ \langle \delta \mu \delta A \rangle & \langle \delta A^2 \rangle \end{pmatrix} \propto \frac{1}{T} \begin{pmatrix} w & 0 \\ 0 & 1/w \end{pmatrix}$$

$\delta \mu = \mu - \mu_0$ and so on.

\Rightarrow SPREAD PROPORTIONAL TO $\frac{1}{\sqrt{T}}$

FURTHER: $\langle \delta \mu^2 \rangle \sim w$
 $\langle \delta A^2 \rangle \sim 1/w$

COMMON PROBLEM IN OPTIMIZING DESIGN

\Rightarrow ACCURACY CAN ONLY BE IMPROVED FOR ONE PARAMETER!

INCREASED BACKGROUND RAISES NOISE LEVEL!

7.2.2. A FREE-FORM SOLUTION

8

$$f(x) = \sum_{j=1}^M a_j \delta(x-x_j)$$

quadratic approximation as before

$$[\vec{\nabla} \vec{\nabla} \chi^2]_{ij} = \frac{\partial^2 \chi^2}{\partial a_i \partial a_j}$$

$$\Rightarrow [\vec{\nabla} \vec{\nabla} \chi^2]_{ij} = \sum_{k=1}^N \frac{2T^2}{\sigma_k^2} R(x_k - x_i) R(x_k - x_j)$$

INVERSE OF $\vec{\nabla} \vec{\nabla} \chi^2 \rightarrow$ COVARIANCE MATRIX

GO OVER TO EIGENVECTORS:

$$\sum_{i=1}^M [\vec{\nabla} \vec{\nabla} \chi^2]_{ij} e_e(x_i) = \lambda_e e_e(x_j)$$

$$e = 1, 2, \dots, M$$

~ OBTAIN NUMERICALLY

ANALYTICAL:

$$\vec{\nabla} \vec{\nabla} \chi^2(x, x') = \frac{2T^2}{\sigma^2} \int R(y-x) R(y-x') dy$$

(FINE SAMPLING, ROUGHLY CONSTANT ERRORS)

AUTO-CORRELATION FUNCTION OF R

$$\vec{\nabla} \vec{\nabla} \chi^2 = \frac{2T^2}{\sigma^2} G(|x-x'|) \quad \text{only depends}$$

on separation between x and x'

\Rightarrow eigenvalue equation

$$\frac{2\tau^2}{\sigma^2} \int G(|x-x'|) e_e(x') dx' = \lambda_e e_e(x)$$

\uparrow CONVOLUTION \Rightarrow GO TO FOURIER SPACE

$$\Rightarrow \frac{2\tau^2}{\sigma^2} \hat{G}(\omega) \hat{e}_e(\omega) = \lambda_e \hat{e}_e(\omega)$$

$$\hat{G}(\omega) = \int G(x) \exp(i\omega x) dx$$

$\hat{G}(\omega)$: real, symmetric

$$\Rightarrow \hat{e}_e(\omega) \propto \delta(\omega - \omega_e) \pm \delta(\omega + \omega_e)$$

$$\xrightarrow{\text{FT}} e_e(x) \propto \sin(\omega_e x) \text{ and } \cos(\omega_e x)$$

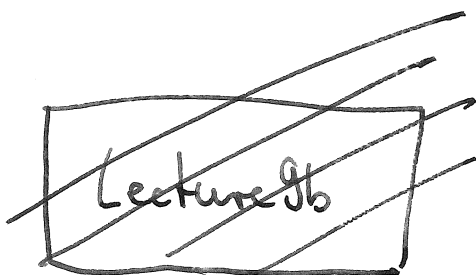
EIGENVALUES: $\lambda_e = \frac{2\tau^2}{\sigma^2} |\hat{R}(\omega_e)|^2$

EXPANSION

$$\Rightarrow f(x) = \sum_e [s_e \sin \omega_e x + c_e \cos \omega_e x]$$

LOOK AT $\text{prob}(\{D_n\} | f(x), I)$ in this space

$$\langle \delta s_e^2 \rangle = \langle \delta c_e^2 \rangle = \frac{2}{\lambda_e}$$



ω, τ same for ω
 $R(\omega)$
 $\omega \rightarrow \frac{1}{4}\omega$ and again

PLOT FROM
BOOK!

ROLE OF BACKGROUND AND OBSERVING TIME

(10)

FOR POISSON DATA $\sigma^2 \sim \# \text{ OF COUNTS}$:

$$\sigma^2 \propto T R_0 (1 + B_f)$$

\uparrow integral of R \uparrow Background

include in $\lambda_e = \frac{2\pi^2}{\sigma^2} |\vec{R}(\omega)|^2$

$$\Rightarrow \lambda_e \propto T$$

\Rightarrow LIKELIHOOD SHARPENS WITH $\frac{1}{\sqrt{T}}$

CHANGING B CONTRIBUTES TO NOISE

CHANGING R_0 (FIG 7.2d, LAST) MAKES BIG CHANGE FOR GAUSSIAN

\Rightarrow LARGER BLUR OF DATA HAS TO BE COMPENSATED BY MANY ORDERS OF MAG. OF OBSERVING TIME!

7.3. CALIBRATION, MODEL SELECTION, BINNING

⑪

NUISANCE PARAMETERS: CAN BE MARGINALIZED

HOW MUCH TIME SHOULD WE SPENT TO CALIBRATE THEM?

BEFORE: BACKGROUND B (FLAT)

IF NOT KNOWN, HOW MUCH TIME SHOULD WE SPENT

$$\text{prob}(\{D_n\} | \{X_{ij}\}, I) = \int \text{prob}(\{D_n\} | \{X_{ij}\}, B, I) \text{prob}(B | I) dB$$

↑
independent
estimate of B

DATA COLLECTED FOR TIME T_D

$$\text{ERRORS} \sim \frac{1}{\sqrt{T_D}}$$

BACKGROUND MEASURE: $\text{prob}(B | I)$ width scales like

$$\frac{1}{\sqrt{T_B}}$$

WHAT RATIO $\frac{T_B}{T_D}$ MAKES INTEGRAL ^{MOST} SHARPLY

PEAKED?

DEPENDS IF B HAS STRUCTURE OR NOT

FOR FLAT, LITTLE USE TO CALIBRATE (SEE EARLIER LECT.)

7.4. INFORMATION GAIN: QUANTIFYING THE WORTH OF AN EXPERIMENT

DEFINE:
$$\mathcal{H} = \int P(x) \log_2 \left[\frac{P(x)}{\pi(x)} \right] dx$$

NEGATIVE ENTROPY \Rightarrow INFORMATION IN POSTERIOR!

POSTERIOR: $\text{prob}(x | D, I)$

RELATIVE TO PRIOR: $\text{prob}(x | I)$

LOGARITHM BASE 2: INFORMATION IS MEASURED IN BITS ∇

EXAMPLE: X JUST TWO STATES

$$\pi(x) = \left(\frac{1}{2}, \frac{1}{2} \right)$$

$$P(x) = (1, 0) \text{ or } (0, 1) \hat{=} \text{FULL KNOWLEDGE}$$

$$\Rightarrow \mathcal{H} = 1 \log_2(2) + 0 \log_2(0) = 1 \text{ bit}$$

FOR FOUR STATES: $\pi(x) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right)$

FULL KNOWLEDGE: $\mathcal{H} = 2 \text{ bits}$

Experiment with DATA D TO MEASURE x

LIKELIHOOD: $\mathcal{L}(D, x) = \text{prob}(D | x, I)$

$$\text{prob}(D | x, I) = \mathcal{L}(D, x) \pi(x)$$

$$\text{EVIDENCE: } Z(D) = \text{prob}(D | I) = \int \mathcal{L}(D, x) \pi(x) dx$$

$$\text{POSTERIOR: } P(x) = \frac{\mathcal{L}(D, x) \pi(x)}{Z(D)}$$

$$\Rightarrow H(D) = \int \frac{\mathcal{L}(D, x) \pi(x)}{Z(D)} \log_2 \left(\frac{\mathcal{L}(D, x)}{Z(D)} \right) dx$$

BEFORE ACQUIRING DATA D

$$\langle H \rangle = \int H(D) Z(D) dD$$

BENEFIT OF EXPERIMENT!

(AMOUNT OF INFORMATION ABOUT x)

8 LEAST SQUARE EXTENSIONS8.1. CONSTRAINTS AND RESTRAINTS

M parameters \vec{X}

set of N data \vec{D}

$$\chi^2 = \sum_{k=1}^N R_k^2$$

$$R_k = \frac{F_k - D_k}{\sigma_k} \quad (\text{RESIDUAL})$$

$$F_k = f(\vec{X}, k) \quad : \text{THEORY}$$

$$D_k : \text{DATA}$$

UNIFORM PRIOR: $\text{prob}(\vec{X} | I) = \text{const}$

uncorrelated Gaussian likelihood:

$$\text{prob}(\vec{D} | \vec{X}, I) = \prod_{k=1}^N \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{R_k^2}{2}\right) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

$$L = \ln[\text{prob}(\vec{X} | \vec{D}, I)] = \text{const.} - \frac{\chi^2}{2}$$

(FOR $N \gg M$; posterior usually irrelevant)

ASSUME GAUSSIAN PRIOR:

(2)

$$\text{prob}(\vec{x} | \mathbf{I}) = \prod_{j=1}^N \frac{1}{\epsilon_j \sqrt{2\pi}} \exp\left(-\frac{(x_j - x_{0j})^2}{2\epsilon_j^2}\right) \propto \exp\left[-\frac{C}{2}\right]$$

$$C = \sum_{j=1}^N \left(\frac{x_j - x_{0j}}{\epsilon_j}\right)^2$$

$$\Rightarrow \text{posterior pdf} : L = \ln(\text{prob}(\vec{x} | \vec{D}, \mathbf{I})) \\ = \text{const.} - \frac{1}{2} [\chi^2 + C]$$

8.2. NOISE SCALING

ASSUME WE HAVE NO ESTIMATE OF σ_k

WHAT CAN BE DONE?

WE COULD DEFINE : $\sigma_k = \sigma \sqrt{D_k}$

AND TREAT σ AS AN UNKNOWN PARAMETER

$$\Rightarrow \text{prob}(\vec{D} | \sigma, \vec{x}, \mathbf{I}) \propto \frac{1}{\sigma^N} \exp\left(-\frac{\chi^2}{2\sigma^2}\right)$$

WITH $\sigma = 1 \Rightarrow \sigma_k = \sqrt{D_k}$ POISSON-LIKE

prior (Jeffrey's) : $\text{prob}(\sigma | \vec{x}, \mathbf{I}) = \text{prob}(\sigma | \mathbf{I}) = \frac{1}{\sigma}$

$$\Rightarrow \text{prob}(\vec{D} | \vec{x}, \mathbf{I}) = \int_0^\infty \text{prob}(\vec{D} | \sigma, \vec{x}, \mathbf{I}) \text{prob}(\sigma | \mathbf{I}) \\ \propto \int_0^\infty \left(\frac{2t}{\chi^2}\right)^{N/2-1} e^{-t} \frac{dt}{\chi^2}$$

with : $t = \frac{\chi^2}{2\sigma^2}$

\hookrightarrow finite integral $\sim (\chi^2)^{-N/2}$

$$\Rightarrow L = \ln[\text{prob}(\vec{x} | \vec{D}, \mathbf{I})] = \text{constant} - \frac{N}{2} \ln \chi^2$$

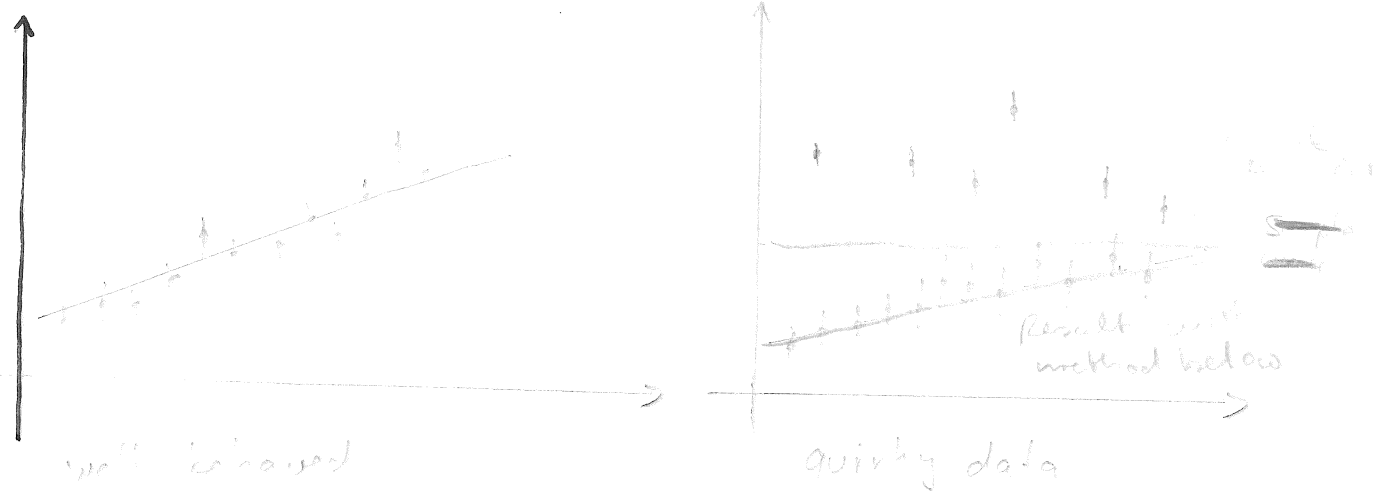
⇒ BEST FIT \vec{x}_0 STILL GIVEN BY χ^2_{\min}

(3)

COVARIANCE:

$$(\vec{\nabla} \vec{\nabla} L)^{-1} = \langle (x_i - x_{0i})(x_j - x_{0j}) \rangle = \underbrace{2 [(\vec{\nabla} \vec{\nabla} \chi^2)^{-1}]_{ij}}_{\text{STANDARD RESULT}} \chi^2_{\min}$$

8.3 OUTLIERS : DEALING WITH ERRATIC DATA



8.3.1. CONSERVATIVE FORMULATION

(if data can not be expunged by hand)

- ASSUME $\{\sigma_n\}$ ARE LOWER ESTIMATE OF NOISE
SINGLE DATUM : D
MISMATCH TO MODEL PREDICTION τ IS GREATER = σ_0

$$\text{prob}(\sigma | \sigma_0, \sigma_1, I) = \frac{1}{\ln(\frac{\sigma_1}{\sigma_0})} \frac{1}{\sigma}$$

IF $\sigma_0 \leq \sigma \leq \sigma_1$

IF UPPER BOUND IS UNKNOWN:

$$\text{prob}(\sigma | \sigma_0, I) = \frac{\sigma_0}{\sigma^2}$$

IF $\sigma \geq \sigma_0$

$$\text{prob}(D|F, \sigma_0, I) = \int_0^{\infty} \text{prob}(D|F, \sigma, I) \text{prob}(\sigma|\sigma_0, I) d\sigma$$

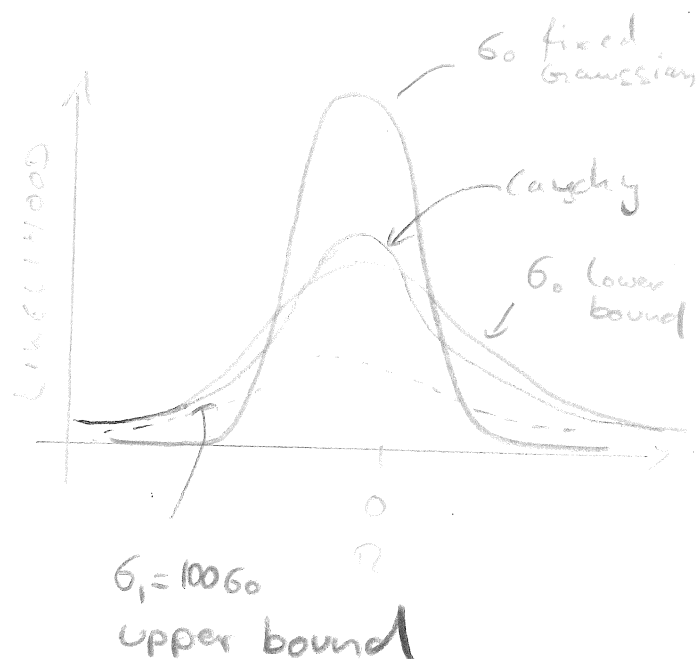
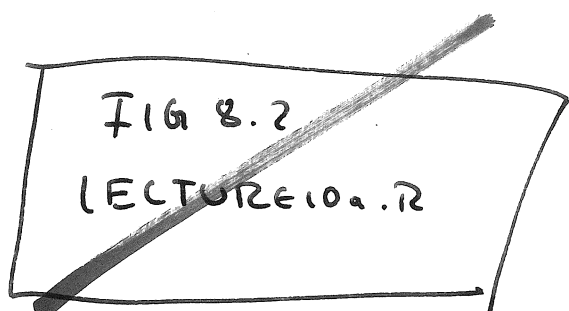
Assume GAUSSIAN FOR $\text{prob}(D|F, \sigma, I)$

$$\text{prob}(D|F, \sigma_0, I) = \frac{\sigma_0}{\sqrt{2\pi}} \int_0^{1/\sigma_0} t e^{-t^2(F-D)^2/2} dt$$

$$\sigma = \frac{1}{t} \quad (d\sigma = -\frac{dt}{t^2})$$

$$\Rightarrow \text{prob}(D|F, \sigma_0, I) = \frac{1}{\sigma_0 \sqrt{2\pi}} \left[\frac{1 - e^{-R^2/2}}{R^2} \right]$$

with $R = \frac{F-D}{\sigma_0}$



MULTIPLE DATA :

$$L = \ln[\text{prob}(\vec{X}|\vec{D}, I)] = \text{const.} + \sum_{k=1}^N \ln \left[\frac{1 - e^{-R_k^2/2}}{R_k^2} \right]$$

(SIMPLE NOISE SCALING "ONLY" INCREASES ERRORS)

$F_n = \mu$; $G_n = 1$ \leftarrow DATA WITH ONE OUTLIER

Lecture 10b

$N=1$
 $N=2$
 $N=4$
 $N=99 \leftarrow$ change x lin

SOLID LINES \leftarrow THE METHOD HERE
DOTTED LINES \leftarrow χ^2

8.3.2 THE GOOD-AND-BAD DATA MODEL

PREVIOUS METHOD: UNCERTAINTIES 50% LARGER

ALLOW FOR TWO POSSIBILITIES: (FOR DATA):

- 1) EVERYTHING IS NORMAL : QUOTED ERROR OK!
- 2) SOMETHING WENT WRONG : SCALE UP NOISE ASSESSMENT!

$$\Rightarrow \text{prob}(\sigma | G_0, \beta, \gamma, I) = \beta \delta(\sigma - \gamma G_0) + (1-\beta) \delta(\sigma - G_0)$$

with $0 \leq \beta \leq 1$ and $\gamma \gg 1$

include in marginalization over Gaussian

$$\Rightarrow \text{prob}(D | F, G_0, \beta, \gamma, I) = \frac{1}{G_0 \sqrt{2\pi}} \left\{ \frac{\beta}{\gamma} \exp\left[-\frac{R^2}{2\gamma^2}\right] + (1-\beta) \exp\left[-\frac{R^2}{2}\right] \right\}$$
$$R = \frac{F - D}{G_0}$$

uniform prior \Rightarrow

$$L = \text{const.} + \sum_{k=1}^N \ln \left[\frac{\beta}{\gamma} e^{-\frac{R_k^2}{2\gamma^2}} + (1-\beta) e^{-\frac{R_k^2}{2}} \right]$$

$\beta \Rightarrow \sigma$: STD. LEAST SQUARES

(EITHER marg. over γ, β , or know a priori)

8.4 BACKGROUND REMOVAL

ASSUME GAUSSIAN NOISE

$$\text{prob}(D|A, B, \sigma, I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(A+B-D)^2}{2\sigma^2}\right]$$

B: Background

A: Signal

D: DATA (SINGLE)

~~RE~~ MARGINALIZE OVER A WITH SUITABLE PRIOR

assume A positive and mean μ $\xrightarrow{\text{Max Ent}}$

$$\text{prob}(A|\mu, I) = \frac{1}{\mu} \left[\exp\left(-\frac{A}{\mu}\right) \right] \quad \text{for } A \geq 0$$

(zero otherwise)

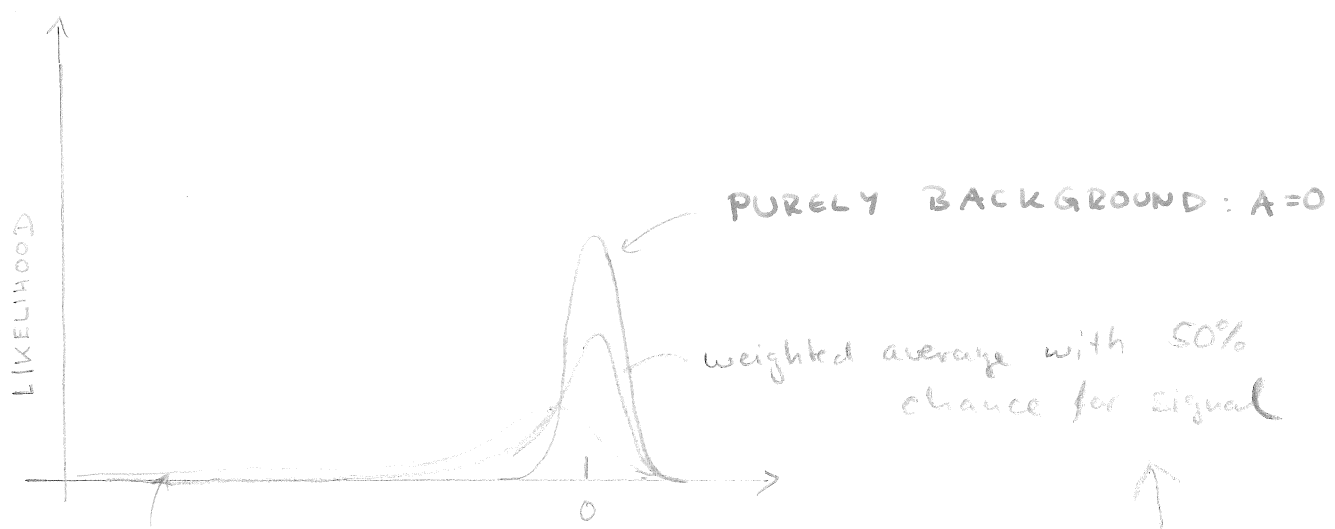
marginalizing

$$\text{prob}(D|B, \mu, \sigma, I) = \frac{1}{\mu\sigma\sqrt{2\pi}} \int_0^{\infty} e^{-A/\mu} e^{-(A+B-D)^2/2\sigma^2} dA$$

$$= \frac{e^{n(R+n/2)}}{2\mu} \left[1 - \text{erf}\left(\frac{n+R}{\sqrt{2}}\right) \right]$$

$$n \equiv \frac{\sigma}{\mu}$$

$$R \equiv \frac{B-D}{\sigma}$$



MARGINAL DISTRIBUTION
 $n^{-1} = \frac{\langle A \rangle}{\sigma} = 10$

$\beta = 0.5$
 $n = 0.1$

INVOLVING GOOD - AND - BAD DATA MODEL

$$0 < \beta < 1$$

$$\text{Prob}(D | B, \beta, \mu, \sigma, I) = \frac{\beta e^{n(R + \pi/2)}}{2\mu} \left[1 - \text{erf} \left(\frac{n+R}{\sqrt{2}} \right) + \frac{(1-\beta) e^{-R^2/2}}{\sigma \sqrt{2\pi}} \right]$$

8.5 CORRELATED NOISE: AVOIDING OVER-COUNTING

UP TO NOW: ASSUMED THAT MEASUREMENTS ARE INDEPENDENT

FOR MULTIVARIATE GAUSSIAN

$$\text{Prob}(\vec{D} | \vec{X}, I) = \frac{1}{\sqrt{(2\pi)^N \det C}} \exp \left[-\frac{1}{2} (\vec{F} - \vec{D})^T C^{-1} (\vec{F} - \vec{D}) \right]$$

with COVARIANCE MATRIX C

$$C_{kk'} = \begin{cases} \sigma_k^2 & \text{for } k=k' \\ 0 & \text{otherwise} \end{cases}$$

IF THIS DOES NOT HOLD:

8

$$\chi^2 = \sum_{k=1}^N \sum_{k'=1}^N (F_k - D_k) [C^{-1}]_{kk'} (F_{k'} - D_{k'})$$

8.5.1 NEAREST-NEIGHBOUR CORRELATIONS

$$\langle (F_k - D_k)(F_{k'} - D_{k'}) \rangle = \begin{cases} \sigma_k^2 & \text{for } k=k' \\ \epsilon \sigma_k \sigma_{k'} & \text{for } |k-k'|=1 \end{cases}$$

with $-1 < \epsilon < 1$: CORRELATION STRENGTH

DOES NOT DEFINE COVARIANCE COMPLETELY:

USE MaxEnt

$$[C^{-1}]_{kk'} = \begin{cases} \Lambda_k & \text{for } k=k' \\ \lambda_k & \text{for } |k-k'|=1 \end{cases}$$

$\{\Lambda_k\}, \{\lambda_k\}$ Lagrange MULTIPLIERS

SOLUTION:

$$\Lambda_k = \begin{cases} \frac{1}{(1-\epsilon^2) \sigma_k^2} & \text{for } k=1 \text{ or } N \\ \frac{1+\epsilon^2}{(1-\epsilon^2) \sigma_k^2} & \text{for } 1 < k < N \end{cases}$$

$$\lambda_k = - \frac{\epsilon}{(1-\epsilon^2) \sigma_k \sigma_{k+1}}$$

$$\Rightarrow C_{kk'} = \sigma_k \sigma_{k'} \epsilon^{|k-k'|}$$

$$\det(\mathbf{C}) = (\sigma_1 \sigma_2 \dots \sigma_N)^2 (1 - \epsilon^2)^{N-1}$$

with uniform prior

$$\Rightarrow L = \text{const.} - \frac{1}{2} \left[(N-1) \ln(1 - \epsilon^2) + \frac{Q}{1 - \epsilon^2} \right]$$

WITH: $Q = \chi^2 + \epsilon \left[c(\chi^2 - \phi) - 2\psi \right]$

$$\chi^2 = \sum_{k=1}^N R_k^2 \quad ; \quad \phi = R_1^2 + R_N^2$$

$$\psi = \sum_{k=1}^{N-1} R_k R_{k+1}$$

(for $\epsilon \rightarrow 0$ std. χ^2)

IF WE THINK THERE IS PROBLEM WITH ERRORS

SCALE: $\gamma \sigma_k$; marginalize over γ

$$L = \text{const.} - \frac{1}{2} \left[N \ln Q - \ln(1 - \epsilon^2) \right]$$

8.5.2. AN ELEMENTARY EXAMPLE

N MEASUREMENT OF QUANTITY μ : $\{x_k\}$

ALL ERROR: σ

~~8.5.2~~

WITHOUT CORRELATION:

$$\mu_0 = \frac{\sum_k x_k}{N}$$

$$\sigma_0 = \frac{\sigma}{\sqrt{N}}$$

NEAREST - NEIGHBOUR CORRELATION:

$$\left. \begin{array}{l} \bar{F}_k = \mu \\ D_k = x_k \\ G_k = \sigma \end{array} \right\} \Rightarrow \mu_0 = \frac{1}{N - \epsilon(N-2)} \left[x_1 + x_N + (1-\epsilon) \sum_{k=2}^{N-1} x_k \right]$$

$$\text{AND } \mu = \mu_0 \pm \sigma \sqrt{\frac{1+\epsilon}{N(1-\epsilon)+2\epsilon}} \approx \mu_0 \pm \frac{\sigma}{\sqrt{N}} \sqrt{\frac{1+\epsilon}{1-\epsilon}}$$

(large N ; $\epsilon < 1$)

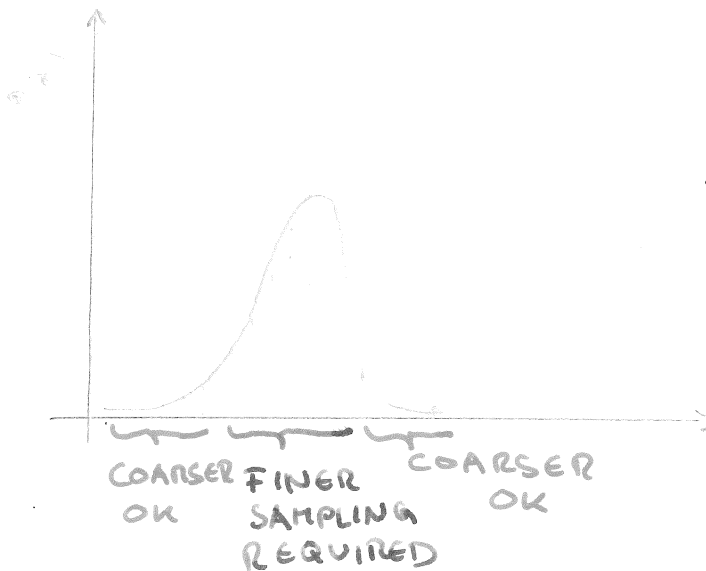
$$\text{IF } \epsilon \rightarrow 1: x_k = x \Rightarrow \mu = x \pm \sigma$$

$$\text{IF } \epsilon \rightarrow -1: \mu \text{ without uncertainty}$$

MONTÉ CARLO MARKOV CHAINMETHODS

LITERATURE: "MARKOV CHAIN MONTE CARLO IN PRACTICE"
GILKS et al., CHAPMAN & HALL/CRC
- THE MORE PARAMETERS, THE LONGER
IT TAKES TO "SCAN" ~~THE~~ POSTERIOR

GRID BASED METHOD



⇒ USUALLY TIME
WASTED IN
TAILS

IF N PARAMETERS, GRID SCALES LIKE T^N

⇒ MONTE CARLO MARKOV CHAIN SAMPLING

POSTERIOR DISTRIBUTION

$$p(\tilde{\theta} | D, I) = \frac{p(D | \tilde{\theta}, I) p(\tilde{\theta}, I)}{\int p(\tilde{\theta}, I) p(D | \tilde{\theta}, I) d\tilde{\theta}}$$

IN GENERAL INTERESTED IN QUANTITIES LIKE
MEAN, VARIANCE, ETC.

$$E[f(\tilde{\theta}) | D, I] \propto \int f(\tilde{\theta}) p(\tilde{\theta} | I) p(D | \tilde{\theta}, I) d\tilde{\theta}$$

(and normalization)

⇒ CALCULATE INTEGRAL

EFFICIENT METHOD: MONTE CARLO
INTEGRATION

SIMPLIFYING ASSUMPTION:

$$E[f(\vec{x})] = \int f(\vec{x}) \pi(\vec{x}) d\vec{x}$$

\vec{x} : N continuous variables

$\pi(\vec{x})$: DISTRIBUTION

MONTÉ CARLO METHOD:

DRAW SAMPLES $\{\vec{x}_t, t = 1, \dots, N_s\}$ from $\pi(\vec{x})$

THEN APPROXIMATE:

$$E[f(\vec{x})] \approx \frac{1}{N_s} \sum_{t=1}^{N_s} f(\vec{x}_t)$$

[N_s LARGE ENOUGH THAT APPROX WORKS]

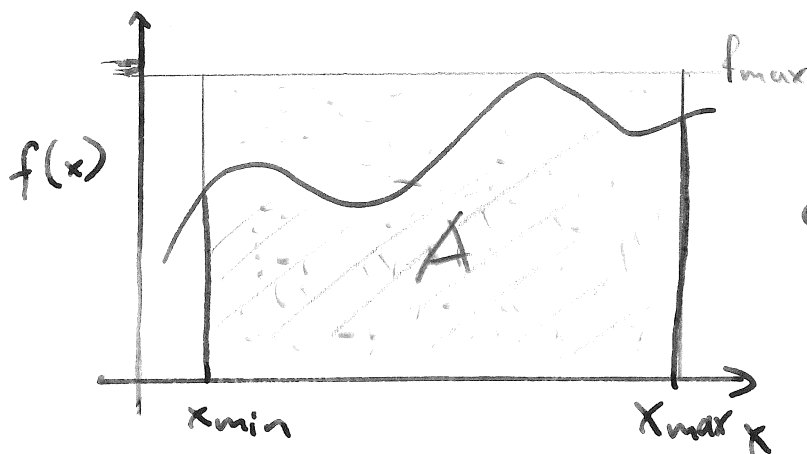
IN GENERAL HARD TO DRAW FROM $\pi(\vec{x})$

BECAUSE $\pi(\vec{x})$ CAN BE ANY DISTRIBUTION

HOWEVER: \vec{x}_t NOT NEEDED TO BE INDEPENDENT

GENERATE IN CORRECT PROPORTIONS ACCORDING TO $\pi(\vec{x})$

INTERLUDE MONTÉ CARLO INTEGRATION:



~~DIVIDE $\{x_{min}, x_{max}\}$~~
~~INTO INTERVALS Δx~~

o GENERATE UNIFORM
RANDOM PAIRS
 (x, y) with
 $x_{min} \leq x \leq x_{max}$
 $0 \leq y \leq f_{max}$

o ACCEPT IF $y \leq f(x)$

o SUM ACCEPTED: N_{acc}

o SUM TOTAL

FOR LARGE SAMPLE N_{TOT}

AREA : $A \approx \frac{N_{acc}}{N_{TOT}} \cdot (x_{max} - x_{min}) \cdot f_{max}$

o VERY EFFICIENT FOR COMPLICATED FUNCTIONS

MARKOV CHAIN WITH $\pi(\vec{x})$ AS LIMITING (STATIONARY) DISTRIBUTION

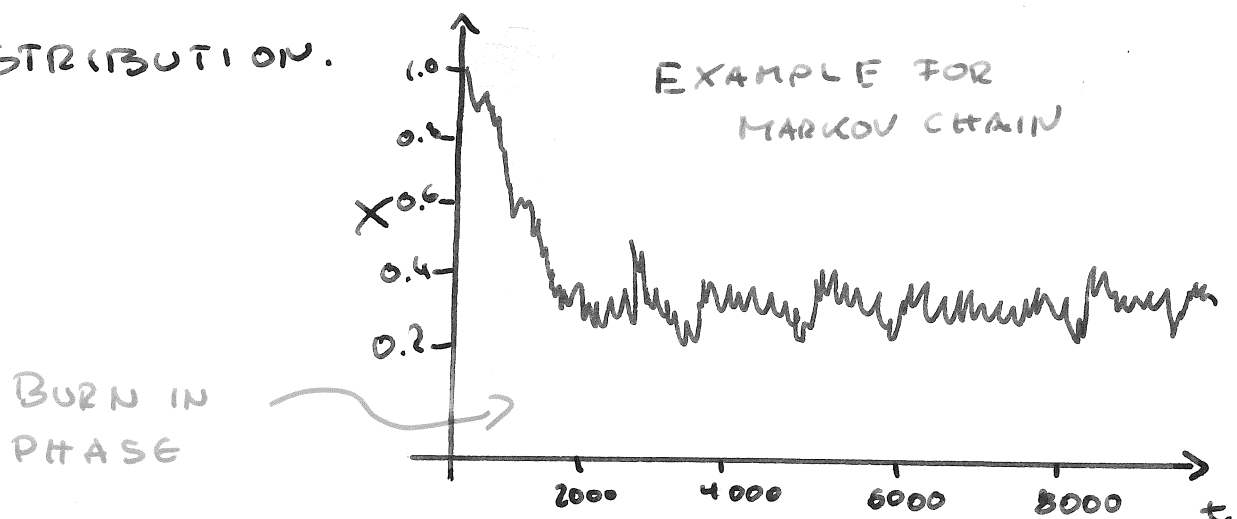
SEQUENCE OF RANDOM VARIABLES $\{\vec{x}_0, \vec{x}_1, \vec{x}_2, \dots\}$

WITH NEXT STATE SAMPLED FROM

$P(\vec{x}_{t+1} | \vec{x}_t) \Rightarrow$ NEXT STATE ONLY DEPENDS ON CURRENT STATE AND NOT ENTIRE HISTORY

\Rightarrow SUCH SEQUENCE IS CALLED : MARKOV CHAIN WITH TRANSITION KERNEL $P(\cdot | \cdot)$

IF PROBABILITY IS WELL BEHAVED, CHAIN WILL "FORGET" ABOUT \vec{x}_0 AND APPROACH A STATIONARY DISTRIBUTION.



THE NEXT STEP IS TO CONSTRUCT A MARKOV

CHAIN WHERE THE STATIONARY DISTRIBUTION

$\phi(\vec{x})$ IS $\pi(\vec{x})$!

ONE POSSIBILITY

METROPOLIS - HASTINGS ALGORITHM

- AT EACH STEP 't' A CANDIDATE POINT \vec{y} IS CHOSEN FROM A PROPOSAL DISTRIBUTION

$$q(\cdot | \vec{x}_t)$$

- EXAMPLE: MULTIVARIATE GAUSSIAN
WITH MEAN \vec{x}_t AND FIXED
COVARIANCE

- THE CANDIDATE POINT IS THEN ACCEPTED
WITH PROBABILITY $\alpha(\vec{x}_t, \vec{y})$, WHERE

$$\alpha(\vec{x}, \vec{y}) = \min \left(1, \frac{\pi(\vec{y}) q(\vec{x} | \vec{y})}{\pi(\vec{x}) q(\vec{y} | \vec{x})} \right)$$

↑ symmetric
for Gaussian!

IF POINT IS ACCEPTED; NEXT STATE

$$\vec{x}_{t+1} = \vec{y}$$

IF REJECTED: $\vec{x}_{t+1} = \vec{x}_t$

METROPOLIS - HASTINGS ALGORITHM

1. INITIALIZE RANDOM \vec{x}_0
2. SAMPLE POINT FROM $q(\cdot | \vec{x}_t)$
3. SAMPLE UNIFORM (0,1) VARIABLE u .
4. IF $u \leq \alpha(\vec{x}_t, \vec{y})$ SET $\vec{x}_{t+1} = \vec{y}$, OTHERWISE
SET $\vec{x}_{t+1} = \vec{x}_t$
5. INCREMENT t AND START AGAIN AT 1.

INTERESTINGLY PROPOSALS $q(\cdot | \cdot)$ CAN HAVE ANY FORM AND STATIONARY DISTRIBUTION WILL STILL BE $\pi(\cdot)$

METROPOLIS ALGORITHM: SYMMETRIC PROPOSALS

$$\Rightarrow q(\vec{y} | \vec{x}) = q(\vec{x} | \vec{y})$$

(here $q(\cdot | \cdot)$ MULTIVARIATE GAUSSIAN) with covariance Σ

$$\Rightarrow \alpha(x, y) = \min\left(1, \frac{\pi(\vec{y})}{\pi(\vec{x})}\right)$$

IMPORTANT: HOW TO CHOOSE COVARIANCE Σ
OF PROPOSAL

- IF Σ TOO SMALL \Rightarrow HIGH ACCEPTANCE RATE
 \Rightarrow "SLOW MIXING" OF CHAIN
- IF Σ TOO LARGE \Rightarrow LOW ACCEPTANCE RATE
 \Rightarrow NO MOVEMENT OF CHAIN

LECTURE 11a.R

FIT STRAIGHT LINE
~~WITH POLYNOMIALS?~~
 WITH MLMC

show χ^2 distribution, histogram, contour
 plot

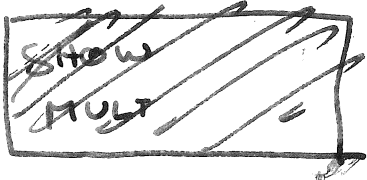
HOW LONG TO RUN A CHAIN?

HAS SEQUENCE CONVERGED AND IS TRULY
 STATIONARY?

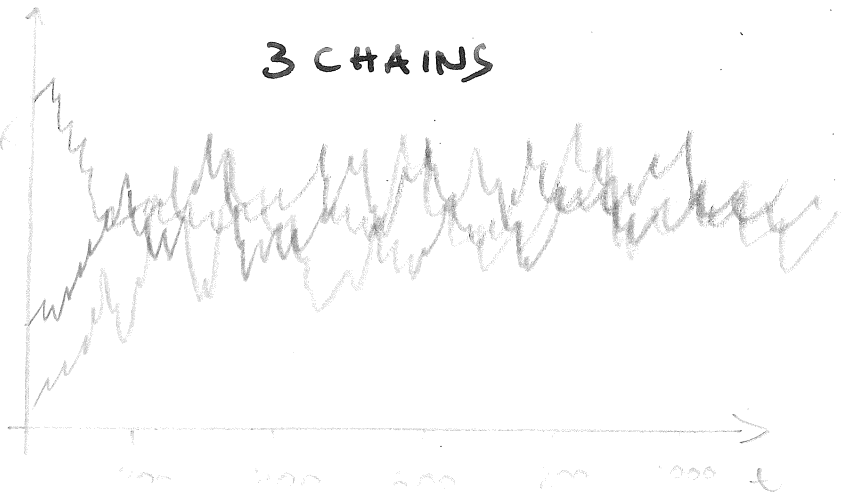
SOLUTION: RUN SEVERAL CHAINS (IN PARALLEL)
 WITH OVER-DISPersed STARTING VALUES

OVER-DISPERSION: FIRST RUN SINGLE CHAIN
 \hookrightarrow DISTRIBUTION OF CHAIN

\Rightarrow USE VARIANCE OF THIS CHAIN
 TO ACHIEVE OVER DISPERSION



3 CHAINS



ASSUME: QUANTITY OF INTEREST ψ
 m PARALLEL SEQUENCES OF LENGTH n

$$\psi_{ij} : j = 1, \dots, n \quad i = 1, \dots, m$$

CALCULATE:
$$B = \frac{n}{m-1} \sum_{i=1}^m (\bar{\psi}_i - \bar{\psi})^2$$

WITH:
$$\bar{\psi}_i = \frac{1}{n} \sum_{j=1}^n \psi_{ij} \quad ; \quad \bar{\psi} = \frac{1}{m} \sum_{i=1}^m \bar{\psi}_i$$

ALSO:
$$W = \frac{1}{m} \sum_{i=1}^m s_i^2$$

WITH:
$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (\psi_{ij} - \bar{\psi}_i)^2$$

W : AVERAGE VARIANCE OF ALL CHAINS

B : VARIANCE OF THE AVERAGES OF CHAINS

(B has factor n based on variance with a sequence of n values, n values are average of n values ψ_{ij})

ESTIMATE OF VARIANCE OF ψ IN TARGET DISTRIBUTION

$$\widehat{\text{var}}(\psi) = \frac{n-1}{n} W + \frac{1}{n} B$$

(OVERESTIMATE)

W: underestimate

(NO TIME FOR INDIVIDUAL CHAINS TO COVER TARGET SPACE)

FOR $n \rightarrow \infty$ BOTH APPROACH TARGET VARIANCE
 $\text{var}(\psi)$

CONVERGENCE TEST:

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\text{var}}(\psi)}{W}}$$

APPROACHES 1 AT CONVERGENCE

Gelman (1996) RECOMMENDS: $\hat{R}^2 - 1 < 0.1$

IMPROVING EFFICIENCY OF SAMPLING

- CHOOSE APPROPRIATE SAMPLING DIRECTIONS AND "STEP" SIZE

EFFICIENT (GAUSSIAN)

PROPOSAL DENSITY SHAPED LIKE TARGET

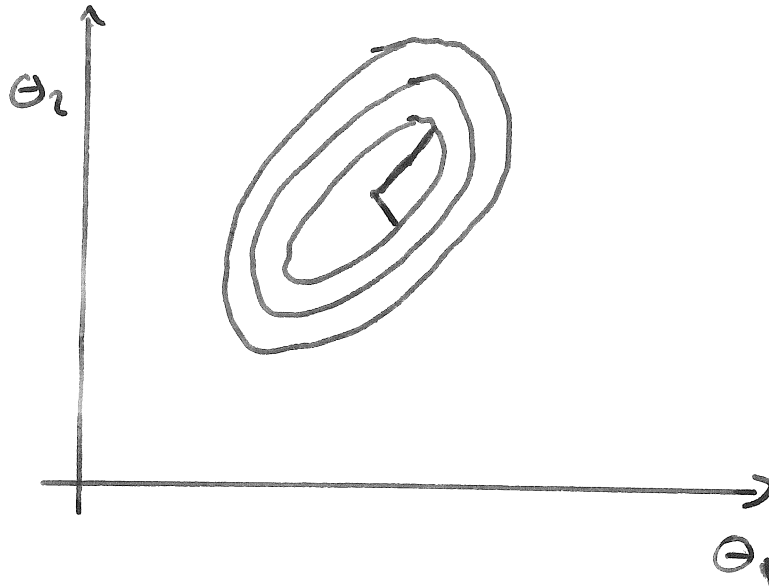
DISTRIBUTION SCALED BY FACTOR $\frac{2.4}{\sqrt{d}}$

(d: # OF PARAMETERS)

TARGET DISTRIBUTION ESTIMATED

FROM EARLY SAMPLES (OR FIRST TEST RUN)
(THROW AWAY)

\Rightarrow CALCULATE COVARIANCE MATRIX



CALCULATE EIGENVALUES AND EIGENVECTORS

CHOOSE GAUSSIAN PROPOSAL DENSITY ALONG
EIGENVECTORS WITH STEPS SCALED BY
EIGENVALUES $\times \sqrt{\lambda^{-1}} \cdot 2.4$

(~~6~~ IF LARGE NON-LINEAR PARAM. DEPENDENCIES
SOMETIMES USEFUL TO TAKE LOGARITHM
OF PARAMETERS)

NO LECTURE NEXT
WEEK