Astrostatistics

modern statistical methods for astronomy

Lecture 2 — 27.1.25

 data (x_i, y_i)

① Where did it come from?

Invert question - if you knew "true" values (η_i, ξ_i), how do you generate the data

↖ latent variables

Posit: $x_i | \xi_i \sim N(\xi_i, \sigma_x^2)$ ↗ $x_i = \xi_i + \epsilon_x, \epsilon_x \sim N(0, \sigma_x^2)$

$y_i | \eta_i \sim N(\eta_i, \sigma_y^2)$

MEASUREMENT ERROR MODEL

② Where do η_i, ξ_i come from?

ⓐ η_i?

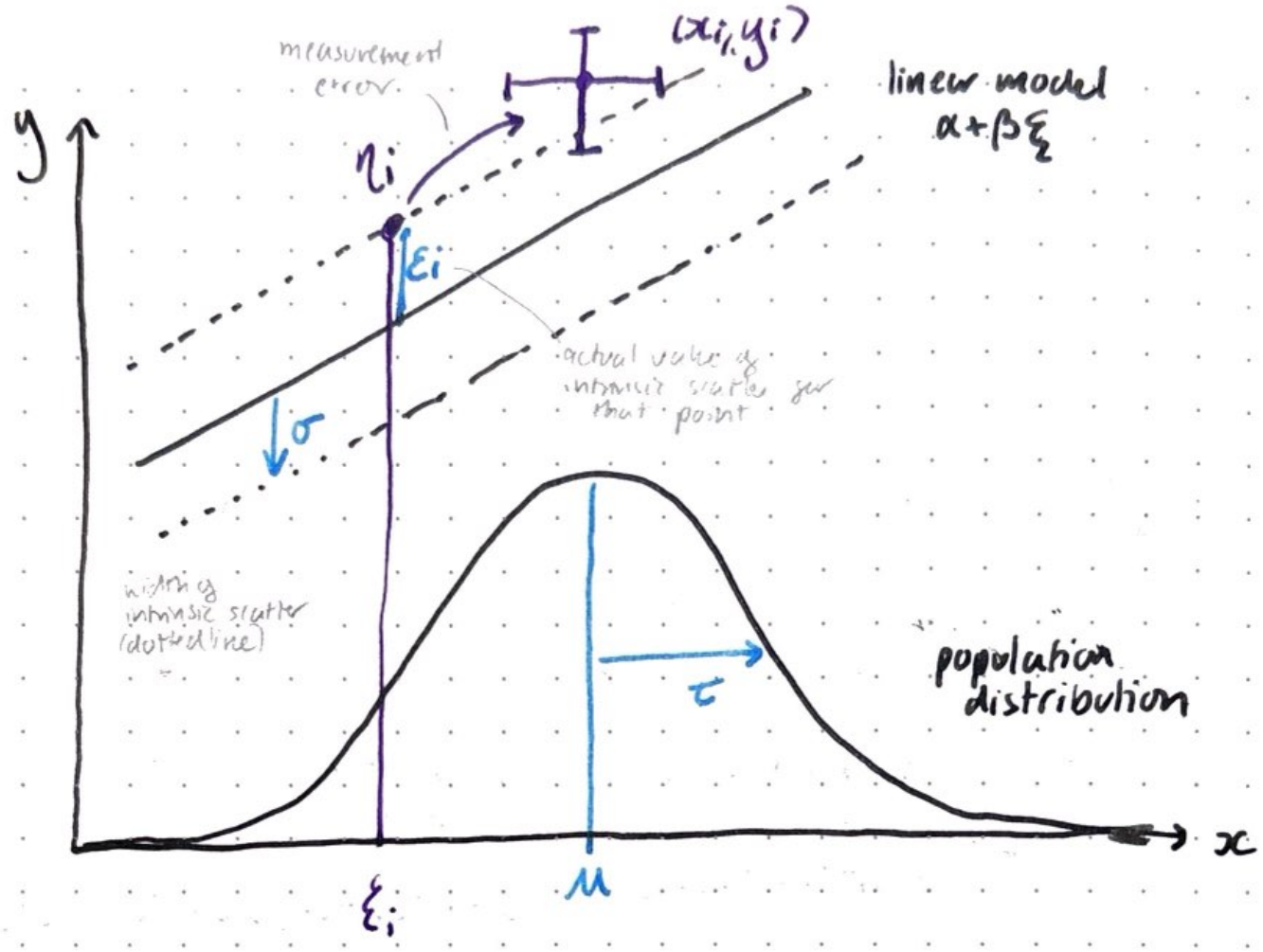
Posit: $\eta_i = \underbrace{\alpha + \beta \xi_i}_{\text{linear model}} + \epsilon_i$ ← intrinsic scatter $\epsilon_i \sim N(0, \sigma^2)$

③ Where does ξ_i come from?

Posit: $\xi_i \sim N(\mu, \tau^2)$

"hyperparameters"

POPULATION DISTRIBUTION MODEL



GENERATIVE MODEL

$\xi \sim N(\mu, \tau^2)$ — Population Distribution

$\eta_i | \xi_i \sim N(\alpha + \beta\xi_i, \sigma^2)$ — Regression Model

$x_i | \xi_i \sim N(\xi, \sigma_x^2)$ } Measurement Error

$y_i | \eta_i \sim N(\eta, \sigma_y^2)$ } Measurement Error

Knowns and Unknowns:

$\underline{\Psi} = (\mu, \tau)$ — Population Distribution Independent Variable "Hyperparameters"

$\underline{\theta} = (\alpha, \beta, \sigma^2)$ — Regression Parameters

(ξ_i, η_i) — Latent (true) Variables ((x,y) without measurement error)

(x_i, y_i) — Observed Data

$(\sigma_{x,i}, \sigma_{y,i})$ — with measurement uncertainties

Lecture 3

29.1.25

FUNDAMENTALS AND NOTATION

(F&B ch. 2-3, Ivezic ch. 3-5)

possible outcome has prob. distribution reflecting uncertainty in outcome

(measurement we take (some number))

Data are realisations of random variables (outcomes of probabilistic experiments / observations)

A random variable X can take on values in some domain (measurable space):

e.g. (discrete) $\mathbb{Z}$, (continuous) $\mathbb{R}$, (multivariate / vector continuous) $\mathbb{R}^N$

Discrete case:

Define the probability distribution over possible outcomes with a probability mass function (PMF)

$$\Pr(X=k) = P_X(k)$$

prob RV takes some value k

e.g. $\Pr(0 \leq X \leq 2) = \sum_{k=0}^{2} P_X(k)$

Normalisation: $\sum_{k=0}^{\infty} P_X(k) = 1$

Continuous case:

ie normally don't work w/ discrete e.g. 10^20 photons → continuous

Define the probability density function (PDF)

$$\Pr(x \leq X \leq x+dx) = P_X(x)\,dx$$

e.g. $\Pr(0 \leq X \leq 2) = \int_0^2 P_X(x)\,dx$

Normalisation: $\int_{-\infty}^{\infty} P_X(x)\,dx = 1$

and the cumulative distribution function (CDF)

$$\Pr(X \leq x_0) = \int_{-\infty}^{x_0} P_x(x)\,dx$$

Multivariate continuous case:

$\vec{x} \in \mathbb{R}^N$ ("member of"), $S \subset \mathbb{R}^N$ ("subset")

$$\Pr(\vec{X} \in S) = \int_S P_{\vec{x}}(\vec{x})\,d^N\vec{x}$$

$d^N\vec{x}$: volume element in $\mathbb{R}^N$

("small x" is values RV X takes on)

Normalisation: $\int_{\mathbb{R}} P_{\vec{x}}(x)\,dx = 1$

When there is no ambiguity, we may simplify notation,

i.e. $P_x(x) \to P(x)$.

("random variable")

Given a distribution $P(x)$ for RV x,

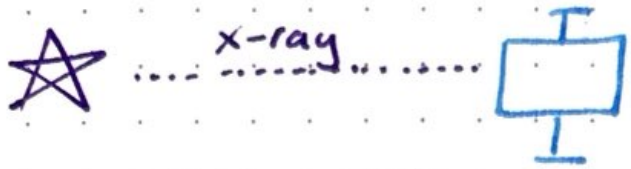
$X \sim P(x)$ means "distributed as", "is drawn from", "draw X from $P(x)$" depending on context

(eg in an algorithm means this)

not "approximation" sign

Astronomers' measurements are produced by physical processes that are inherently probabilistic (truly random processes thanks to quantum mechanics).

Example 1: Photometry / Photon counting



Chandra x-ray observatory

$$\text{rate} = \frac{\text{photons arriving}}{\text{time}} = r$$

$$\Pr(k \text{ photons in time } T) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{where } \lambda = rT$$

$$P(k) = \text{Poisson}(k|\lambda)$$

can also be written.

$$k \sim \text{Poisson}(\lambda)$$

Gaussian/Normal

for $k \to$ large, $P(k) \simeq N(k|\lambda,\lambda)$ a limit theorem

Define Gaussian Random Variable

$$X \sim N(\mu, \sigma^2) \quad , \quad X \in \mathbb{R}$$

PDF: $$P(x) = N(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$$

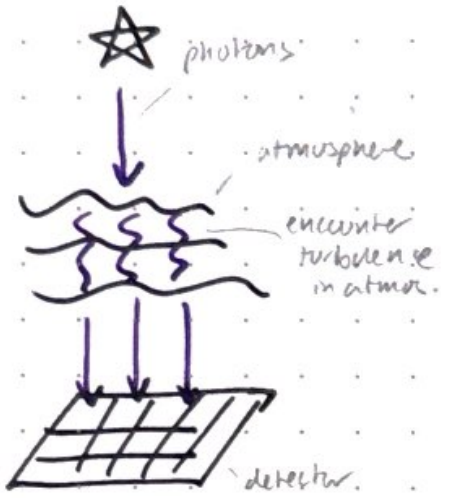
Generalise to Multivariate Gaussian Vector

$$\vec{v} \in \mathbb{R}^N \quad (\text{eg. } N=2, \ \vec{v} = \begin{pmatrix} x \\ y \end{pmatrix})$$

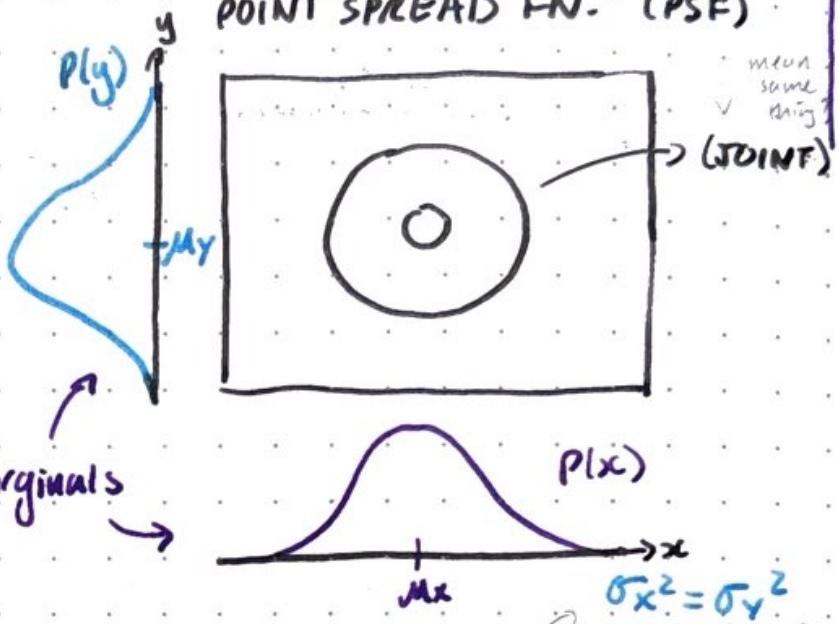
PDF: $$P(\vec{v}) = N(\vec{v}|\vec{\mu}, \underset{\sim}{\Sigma}) = |2\pi\underset{\sim}{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\vec{v}-\vec{\mu})^T \underset{\sim}{\Sigma}^{-1}(\vec{v}-\vec{\mu})}$$

$\vec{\mu} \in \mathbb{R}^N$, $\underset{\sim}{\Sigma}$ is $N \times N$ symmetric pos. def. matrix.

Example: Astrometry



POINT SPREAD FN. (PSF)



$\sigma_x^2 = \sigma_y^2$ no preferred direction

Joint distribution $P(x,y) = \mathcal{N}\left(\begin{pmatrix} x \\ y \end{pmatrix} \middle| \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}\right)$

simplifying ↓

$= \mathcal{N}(x|\mu_x, \sigma_x^2)\,\mathcal{N}(y|\mu_y, \sigma_y^2)$

Law of total probability (LTP):

$$P(x) = \int_{-\infty}^{\infty} P(x,y)\,dy = \mathcal{N}(x|\mu_x, \sigma_x^2)$$

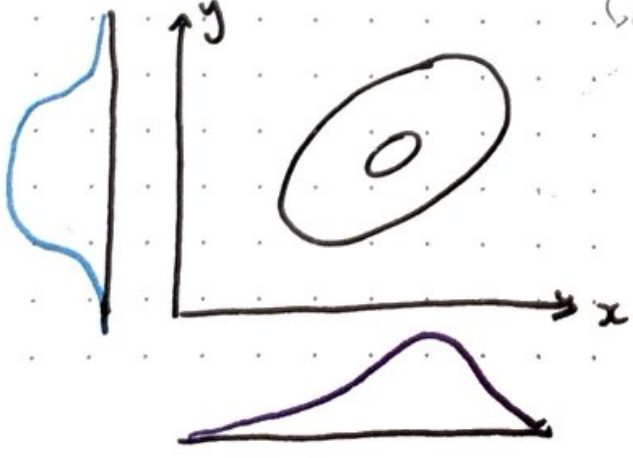
$$P(y) = \int_{-\infty}^{\infty} P(x,y)\,dx = \mathcal{N}(y|\mu_y, \sigma_y^2)$$

In this case, $P(x,y) = P(x)P(y)$, X, Y indep. RVs.

In general, this is not true:

$P(x,y) = \mathcal{N}\left(\begin{pmatrix} x \\ y \end{pmatrix} \middle| \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$ $\quad \rho > 0$

introduces slant



$P(x,y) \neq P(x)P(y)$

marginals same as above

Conditional Probability

lower dim distributes obtained by integrating out other variables

marginals

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

joint — conditionals

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)P(y)}{P(x)}$$

BAYES' THEOREM

Note that LTP: $P(x) = \int P(x,y)\,dy = \int P(x|y)P(y)\,dy$

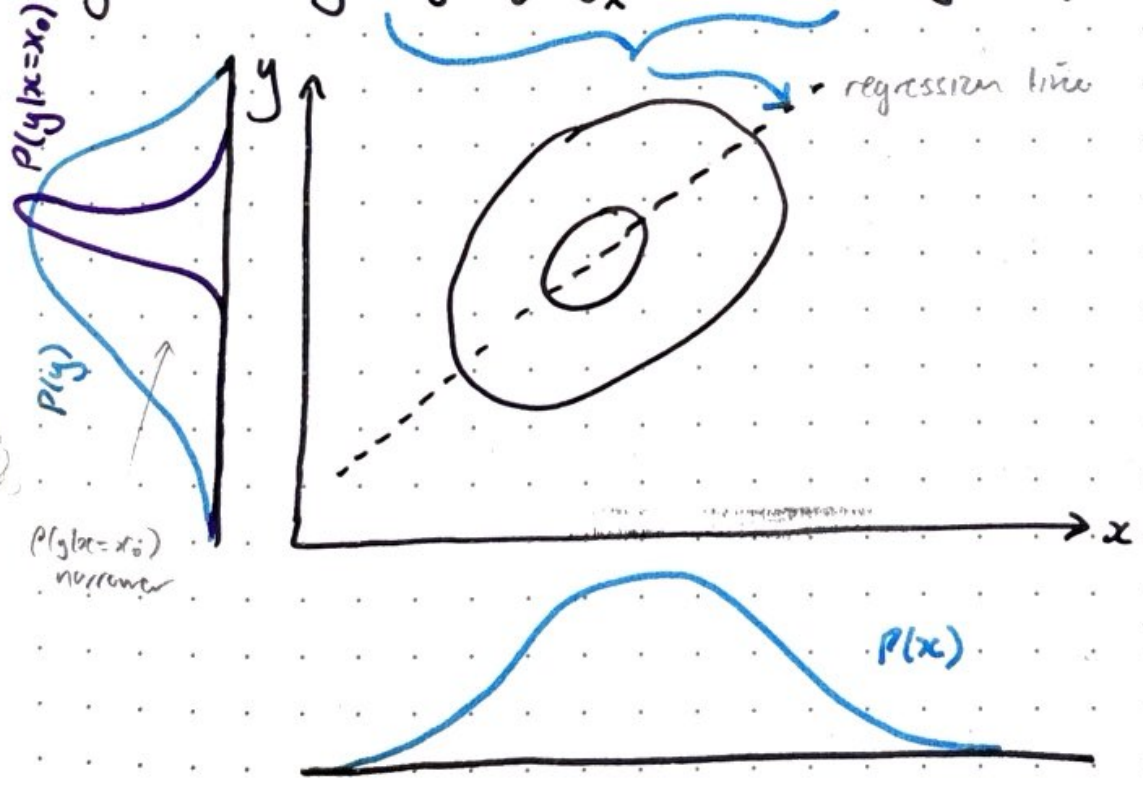
$$P(y|x) = \frac{P(x|y)P(y)}{\int P(x|y)P(y)\,dy}$$

In the bivariate case: Gaussian

$$P(y|x) = N\left(y \mid \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x),\ \sigma_y^2(1-\rho^2)\right)$$

if i know x, i can predict y as it is bivariate gaussian b/c i know x, variance in y decreases by factor $(1-\rho^2)$

homoscedasticity: a property of gaussians (learn this)



Observations x_0 gives you information to predict y

$p(y|x_0) \neq p(y)$, $\rho > 0$.

Lecture 4

31.1.25

MOMENTS & SUMMARIES OF PROB. DISTRIBUTIONS

$X \sim P(x)$ -> X is RV drawn from P(x)

Expect values

$$\mathbb{E}[X] = \int x\, P(x)\, dx$$

$$\mathbb{E}[f(x)] = \int f(x)\, P(x)\, dx$$

Variance

$$\text{Var}(X) = \mathbb{E}[(x - \mathbb{E}(x))^2]$$

$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

pf: $\mathbb{E}(\mathbb{E}(x)) = \mathbb{E}(x)$

Multivariate (X, Y):

Covariance

$$\text{Cov}(X,Y) = \mathbb{E}\left((X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right)$$

$(= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$?)

PROPERTIES: $\text{cov}(X,X) = \text{Var}(X)$, $\text{cov}(X,Y) = \text{cov}(Y,X)$, $\text{cov}(aX, bY) = ab\,\text{cov}(X,Y)$, $\text{cov}(X+a, Y+b) = \text{cov}(X,Y)$

$\text{Cov}(aX+bY, cW+dV) = ac\,\text{cov}(X,W) + ad\,\text{cov}(X,V) + bc\,\text{cov}(Y,W) + bd\,\text{cov}(Y,V)$

Multivariate Normal:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho \\ \sigma_X \sigma_Y \rho & \sigma_Y^2 \end{pmatrix} \right) \qquad -1 < \rho < 1$$

* $X \perp Y \rightarrow \text{Cov}(X,Y) = 0$

* general case $\text{Cov}(X,Y) = \sigma_X \sigma_Y \rho$

correlation $= \rho = \frac{\text{cov}}{\text{scary normality covariance}}$

$\perp$ denotes independence of RVs

Useful Properties:

Collection of random variables

$\{X_1, \ldots, X_N\}, \{Y_1, \ldots, Y_N\}$ <- set of X & Y measurements labelled 1 to N

Def. $S = \sum_{i=1}^{N} a_i X_i$, $T = \sum_{j=1}^{M} b_j Y_j$, $a_i, b_j \in \mathbb{R}$

Expectations are linear → $\mathbb{E}[S] = \sum_{i=1}^{N} a_i \mathbb{E}[X_i]$

expectation of sum is sum of expectation

Bilinearity of covariance: (covariance sum is sum of cov (bilinear in eqn))

$$\mathrm{Cov}(S,T) = \mathrm{Cov}\left(\sum_{i=1}^{N} a_i X_i, \sum_{j=1}^{M} b_j Y_j\right) = \sum_i \sum_j a_i b_j \mathrm{Cov}(X_i, Y_j)$$

$$\mathrm{Var}(S) = \mathrm{Cov}(S,S) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{Cov}(X_i, X_j)$$

(N terms) (N(N-1) terms)

∴ useful for summing multiple sources of error.

↳ variance of a sum of RV is covariance of sum w/ itself

just this term if no errors correlated

some errors correlated

IMPORTANT!

variance of sum is sum of covariances!

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{Cov}(X_i, X_j)$$

$$\left(= \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) + \sum_i \mathrm{Var}(X_i)\right)$$

e.g. $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X,Y)$

TRANSFORMATIONS OF RVs

$X \sim P(x)$ ← invertible & differentiable transformation (fn)

$Y = \phi(X)$ what is $P(y)$?

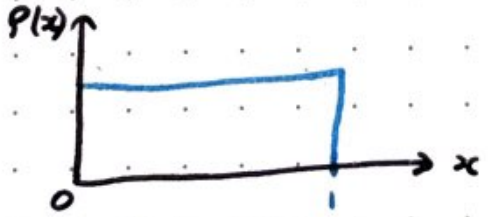
$x = \phi^{-1}(y)$

$$P(y) = P(x)\left|\frac{dx}{dy}\right| = P(\phi^{-1}(y))\left|\frac{d\phi^{-1}(y)}{dy}\right|$$ ← jacobian

(e.g. important if $\underline{X} \sim N(\underline{\mu}, \underline{\underline{\Sigma}})$ then $\underline{Y} = \underline{\underline{A}}\underline{X} + \underline{b} \sim N(\underline{b} + \underline{\underline{A}}\underline{\mu}, \underline{\underline{A}}\underline{\underline{\Sigma}}\underline{\underline{A}}^T)$.
(uses: $\left|\frac{d\underline{y}}{d\underline{x}}\right| = |\underline{\underline{A}}| = \det \underline{\underline{A}}$ & properties of determinants in appendix)

Example:

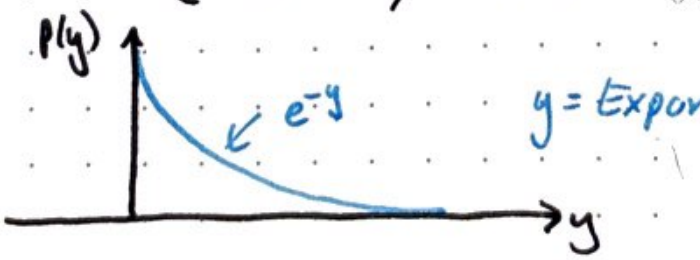
$X \sim \mathcal{U}(0,1)$



$Y = -\ln X$

$X = e^{-Y}$ → $P(y) = \begin{cases} e^{-y}, & 0 \leq y < \infty \\ 0, & \text{o/w} \end{cases}$ otherwise?!

$\frac{dx}{dy} = -e^{-y}$

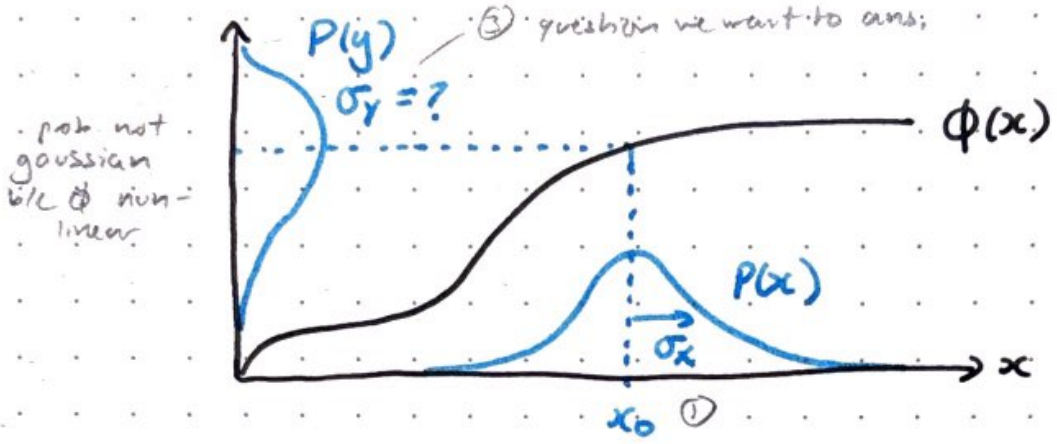


$y = \mathrm{Expon}(1)$

Propagation of error

can think of this as transformation of RV but with variances

Measurement x_0 w/ error variance σ_x^2, error bar σ_x



$$y = \phi(x) = \phi(x_0) + \left.\frac{d\phi}{dx}\right|_{x=x_0} (x - x_0) + \dots$$

$$\mathrm{var}(y) \simeq \left.\frac{d\phi}{dx}\right|^2_{x=x_0} \mathbb{E}\left((x-x_0)^2\right)$$

$$\sigma_y^2 \simeq \left.\frac{d\phi}{dx}\right|^2_{x=x_0} \sigma_x^2$$

ignore this??

$\mathrm{var}(a+X) = \mathrm{var}(X)$?
$\mathrm{var}(bX) = b^2 \mathrm{var}(X)$

$\mathrm{var}(y) = \left|\frac{d\phi}{dx}\right|^2_{x_0} \mathrm{var}(x - x_0)$

$= \left|\frac{d\phi}{dx}\right|^2_{x_0} \mathrm{var}(x)$

Example:

Astronomers' flux f, σ_f

magnitude $m = -2.5 \log_{10}(f) + \text{const.}$

$$\sigma_m \simeq \left(\frac{2.5}{\ln 10}\right) \frac{\sigma_f}{f} \simeq \frac{\sigma_f}{f}$$

$\log_{10} x = \frac{\ln x}{\ln 10}$

astronomers approx $\frac{2.5}{\ln 10} \approx 1$ good enough

goodness of approximation: taylor exp. depends on how linear transformation is in the neighbourhood where we have uncertainty
(will break down if uncertainty large enough or transf. non-linear enough. (see slides))

MULTIVARIATE: $\rightarrow Y = g(X_1, \dots, X_N) \simeq g(X^0) + \sum_i \frac{dg}{dx_i}(X_i - X_i^0)$ linear taylor exp.

$\rightarrow$ use rule $\mathrm{Var}(aX + bY) = a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$

$\rightarrow \underbrace{\mathrm{Var}(Y)}_{\sigma_Y^2} \simeq \sum_i \left(\frac{dg}{dx_i}\right)^2 \underbrace{\mathrm{Var}(X_i)}_{\sigma_X^2} + \sum_{i \neq j} \left(\frac{dg}{dx_i}\right)\left(\frac{dg}{dx_j}\right) \mathrm{Cov}(X_i, X_j)$

$\rightarrow (\rho\, \sigma_{x_i} \sigma_{x_j})$ in terms of correlation

e.g. see wiki

Multivariate case:

→ known Var/Cov

just learn this?

$X_{obs} = (X_1, \ldots, X_N)$, $Y = g(X_1, \ldots, X_N)$

$$\sigma_Y^2 \simeq \sum_{i=1}^{N} \left(\frac{dg}{dx_i}\right)^2 \Big|_{x_{obs}} \mathrm{Var}(X_i) + \sum_{i \neq j} \left(\frac{dg}{dx_i}\right)\left(\frac{dg}{dx_j}\right)\Big|_{x_{obs}} \mathrm{Cov}(X_i, X_j)$$

$$\sigma_Y^2 \simeq \sum_i \left(\frac{dg}{dx_i}\right)^2 \sigma_{x_i}^2 + \sum_{i \neq j} \left(\frac{dg}{dx_i}\right)\left(\frac{dg}{dx_j}\right) \mathrm{cov}(x_i, x_j)$$

important term allowing for correlated errors

iid SEQUENCES OF RVs

— why care? e.g. in experiment do multiple observations, often model them as iid sequence (but not always!)

iid = independent & identically distributed (e.g. repeated measurement)

Suppose we have a sequence of indexed RVs:

$$X_1, X_2, X_3, \ldots, X_N \overset{iid}{\sim} P(X)$$

this means RVs are independent! ($\Rightarrow \mathrm{Cov}(X_i, X_j) = 0$)

identically distrib. →

$$X_i \sim P(X), \quad i \neq j \quad X_i \perp X_j$$

← indep (+ correspondent condition?)

$$\text{Joint distribution} \sim P(X_1, X_2, X_3, \ldots, X_N) = \prod_{i=1}^{N} P(X_i)$$

Not example: $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix}\right) \quad \rho \neq 0$

not indep b/c correlated

May $P(X_1) = \int P(X_1, X_2)\, dX_2 = N(X_1 | \mu, \sigma^2)$, $P(X_2) = N(X_2 | \mu, \sigma^2)$

(can marginalise normal analytically)

X_1, X_2 identically distributed but not independent (or as last lecture $P(X_2|X_1) \neq P(X_2)$.

basically just pick out σ_X, μ_X i think

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)}\sqrt{\mathrm{Var}(X_2)}} = \rho \neq 0.$$

show not independent if:
* $P(a,b) = P(a)P(b)$
or showing
* Corr = 0 (equivalently Cov = 0)

Limit theorems for iid RVs

(see book for technicalities)

① LAW OF LARGE NUMBERS (LLN)

$X_1, \ldots, X_N \overset{iid}{\sim} P(X)$

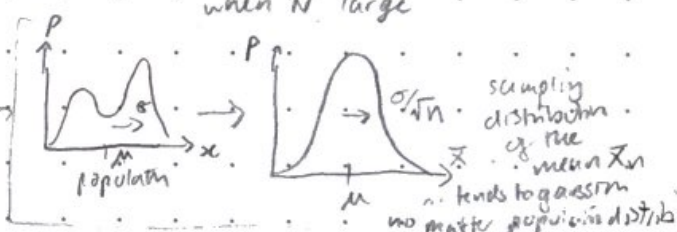
$\mu = \mathbb{E}[X_i] = \int x P(x) dx$ → {theoretical model / population mean}

Def: Sample mean $\bar{X}_N \equiv \frac{1}{N} \sum_{i=1}^{N} X_i$

as $N \to$ large, $\bar{X}_N \to \mu$

i.e. sample mean gets good approx. to population mean when N large

② CENTRAL LIMIT THEOREM



sampling distribution of the mean $\bar{X}_N$ tends to gaussian no matter population distrib

$X_1, \ldots, X_N \overset{iid}{\sim} P(X)$, $\sigma^2 = \mathrm{Var}(X_i)$ finite

note: this is a general result, we haven't assumed $P(X)$ is gaussian, ... is finite

As $N \to$ large, $\frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \to N(0,1)$ (approaches gaussian) — unit gaussian

i.e. $\bar{X}_N \to N(\mu, \sigma^2/N)$ ← error in the mean $= \sigma/\sqrt{N}$

STATISTICAL ESTIMATORS

Probability model $P_\theta(x)$ or $P(x|\theta)$ (data, RV; parameters)

Suppose data are $D = (X_1, \ldots, X_N)$, $X_1, \ldots, X_N \overset{iid}{\sim} P_\theta(x)$

data are ... drawn from prob. distribution

Estimate θ from $\vec{D}$ by constructing an estimator $\hat{\theta}(D)$ (fn. of D) that yields a "good" value for the estimand θ.

thing you are estimating

Lecture 5

3.2.25

CMB → radiation from ~ 300,000 years after big bang
fluctuations in plasma temp of early universe →  map 
planck is a modern map of these fluctuations

distrib of temperatures (at each pixel) measured is "the most perfect gaussian in nature"
an example → how might we find the mean of this distribution?

→ "point estimates" ? see below

STATISTICAL ESTIMATORS CONT.

(that notation important!)

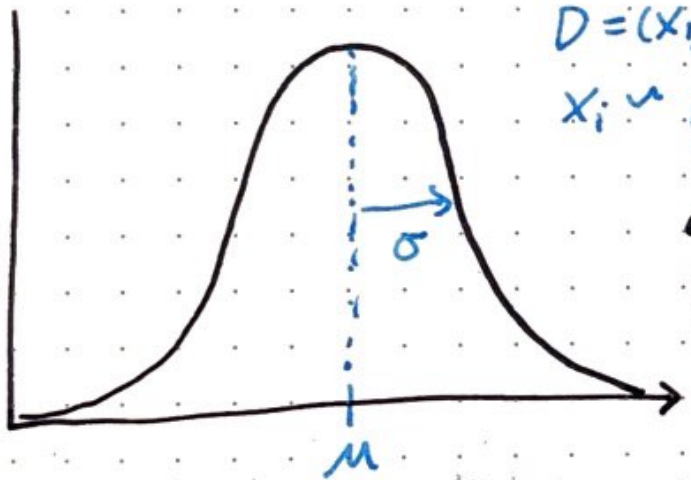
Construct a "good" estimator $\hat{\theta}(D) = \hat{\theta}(\vec{x})$ for estimand θ.

$\hat{\theta}$ ↓ estimate from data θ ↘ the exact true value

Prob model $P_\theta(x)$ or $P(x|\theta)$

$D = (x_1, \dots, x_N)$ $\overset{iid}{\sim} P(x|\theta)$

$x_i \sim P_\mu(x)$, $\mu = \mathbb{E}_P(x)$



← (CMB Example)

eg. estimate μ from temp. measurement in each pixel x_i $i = 1, \dots, N$
Estimand $\mu = \mathbb{E}(x_i)$
Estimator $\hat{\mu} = g(\vec{x})$

Can be many possible estimators for a particular estimand

Possible Estimators:

how do we choose?

① Sample mean $\frac{1}{N}\sum_{i=1}^{N} x_i$ unbiased

② Take $K < N$, $\frac{1}{K}\sum_{i=1}^{K} x_i$ truncated sample mean
sample mean of 1st K of N points
unbiased, inconsistent
only take K samples so closest in size of sample N.

③ $\frac{1}{N-1}\sum_{i=1}^{N} x_i$ not unbiased but asymptotically unbiased

④ Midrange $= \frac{1}{2}[\max(\vec{x}) + \min(\vec{x})]$

⑤ Median

⑥ Bin → Histogram → mode 

⑦ Just report 3 Kelvin

How do you evaluate different possible estimators for a particular estimand?

estimand: μ — estimator: $\hat{\mu}(\vec{x})$

Criteria for Estimators $\hat{\theta}$ for θ

if $\hat{\theta} = \hat{\theta}(\vec{x})$ is estimator for θ,

UNBIASEDNESS: $\hat{\theta}$ is unbiased estimate for θ if

$$\mathbb{E}_p[\hat{\theta}] = \theta,$$ — definition, eg ①, ②

bias — $$b(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

expectation is over repeated experiments sampling the full data distribution $P(\vec{x}|\theta)$

Imagine you did J experiments with fixed sample size N, $j = 1, \dots, J$ — index experiments 1 to J

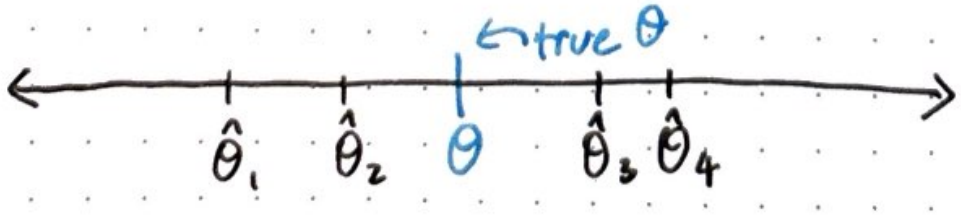
$$\rightarrow \vec{x}_j = (x_{1,j}, \dots, x_{N,j})$$

j experiments, datasets $\vec{x}_j$

$$\vec{x}_1, \vec{x}_2, \dots, \vec{x}_J$$

$$\hat{\theta}_1 = \hat{\theta}(\vec{x}_1) \quad \hat{\theta}_2 = \hat{\theta}(\vec{x}_2) \quad \hat{\theta}_J = \hat{\theta}(\vec{x}_J)$$

Unbiased as $J \to \infty$; $\frac{1}{J}\sum_{j=1}^{J} \hat{\theta}_j = \mathbb{E}[\hat{\theta}] = \theta$



as increase no. of experiments, avg tends to θ

N.B. You only really did one experiment!

ASYMPTOTICALLY UNBIASED: $\left[\mathbb{E}[\hat{\theta}(\vec{x})] \to \theta \text{ as } N \to \infty\right]$

CONSISTENCY: As you gather more data (sample size $N \to$ large), $\hat{\theta}$ converges to θ,

$$\left[\forall \varepsilon > 0,\ Pr(|\hat{\theta} - \theta| > \varepsilon) \to 0 \text{ as } N \to \infty\right]$$

EFFICIENCY: Smallest mean squared error

$$\left[MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta}(x) - \theta)^2]\right]$$

(for unbiased → smallest variance)

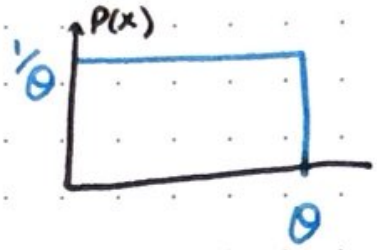
$$\left[\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\ &\dots \\ &= Var(\hat{\theta}) + Bias^2(\hat{\theta}) \end{aligned}\right]$$

BIAS-VARIANCE TRADE OFF: Sometimes the most efficient estimator is biased and the unbiased estimator is not the most efficient.

MINIMUM VARIANCE UNBIASED ESTIMATORS (MVUE):

Caveat: Unbiased estimators can be wrong (obviously wrong)

Ex. $X_i \overset{iid}{\sim} U(0,\theta)$ $\quad i = 1,\dots,N$



Estimate θ?
Unbiased Estimator?

Suppose we take sample mean:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\mathbb{E}[\bar{X}] = \theta/2$$

$\hat{\theta}_u = 2\bar{X}$ unbiased estimator

unbiased

since $\mathbb{E}[\hat{\theta}_u] = 2(\theta/2)$

Suppose instead

always.

$\hat{\theta}_M = \max(X_1,\dots,X_N) < \theta$ biased!

generate samples → RNG $\vec{x} = (0.32, 0.46, 0.97, 2.77, 7.06)$

unbiased estimator → $\hat{\theta}_\mu = 4.63$ unbiased but clearly wrong

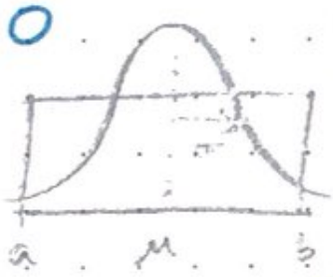
biased estimator → $\hat{\theta}_M = 7.06$ biased but much more sensible

Unbiasedness is not a property of any single experiment, only a property of averaging over all experiments that did not happen.

Ex. Estimator properties depend on data process $P(x|\theta)$, $N=10$

(choose a, b s.t. $\mu = \frac{(a+b)}{2}$, $Var(x_i) = \frac{(b-a)^2}{12} = \sigma^2$)

Estimator / Variance	Gaussian $X_1, \dots, X_N \overset{iid}{\sim} N(\mu, \sigma^2=1)$	Uniform $X_1, \dots, X_N \overset{iid}{\sim} U(a,b)$
Sample Mean $\bar{X}$	$\sigma^2/N = 0.10$	$\sigma^2/N = 0.10$
Midrange $\frac{\min(\vec{x})+\max(\vec{x})}{2}$	$\frac{\pi^2\sigma^2}{24 \ln N} = 0.18$	$\frac{6\sigma^2}{(N+2)(N+1)} \approx \frac{6\sigma^2}{N^2} = 0.06$



midrange less efficient (gaussian) midrange more efficient (uniform)

both sample mean & midrange are unbiased, but midrange st.deviation V goes down as $1/N$ whereas much slower for mean which goes down as $1/\sqrt{N}$ (for uniform) see slides

$(\rightarrow b^3 - a^3 = (b-a)(b^2 + ab + a^2))$

$\rightarrow$ midrange uniform > gaussian, choose a, b s.t. variance σ^2

LIKELIHOOD-BASED INFERENCE

Probability model $P(\vec{D}|\theta)$ we call "sampling distribution" → sampling dist.

Probability distribution for possible / potential datasets $\vec{D}$, for a given parameter value θ.

We observe $\vec{D}_{obs}$: $P(\vec{D} = \vec{D}_{obs}|\theta) = L(\theta)$

↑ Likelihood Function

Sampling dist. probability dist. over all possible outcomes of $\vec{D}$ for a given θ ("before it is observed") N.B. $\int f(\vec{D}|\theta) d\vec{D} = 1$

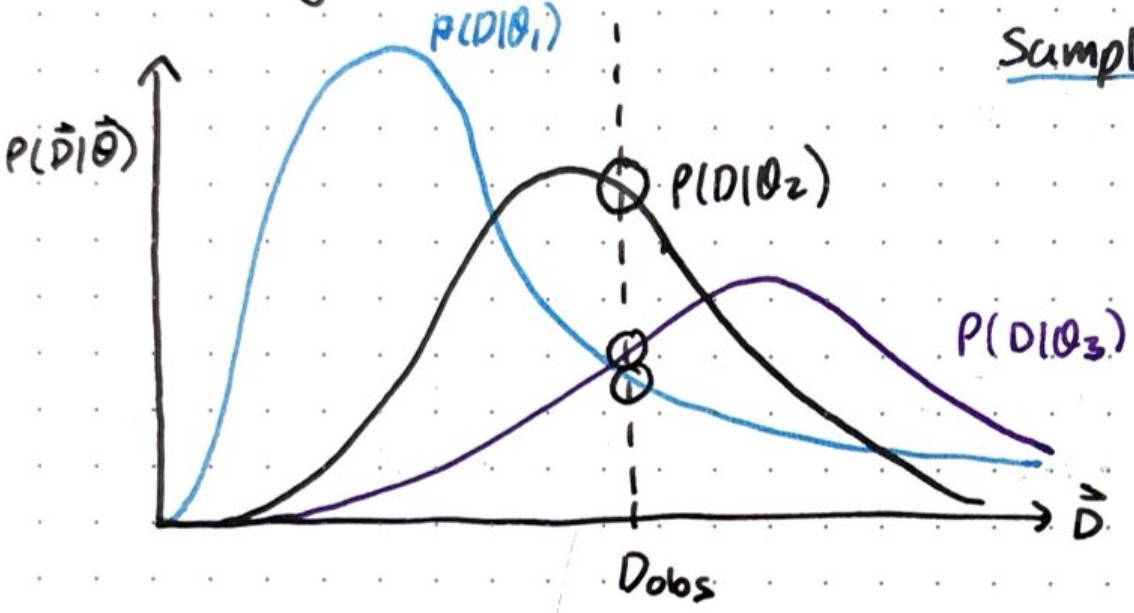
another technique of point estimation: $\hat{\theta}_{MLE}$
"chance of given set of observations occurring"

Lecture 6

5.2.25

LIKELIHOOD-BASED INFERENCE CONT.

Probability model $\vec{D} \sim P(\vec{D}|\vec{\theta})$, $\int P(\vec{D}|\vec{\theta})\,d\vec{D} = 1$

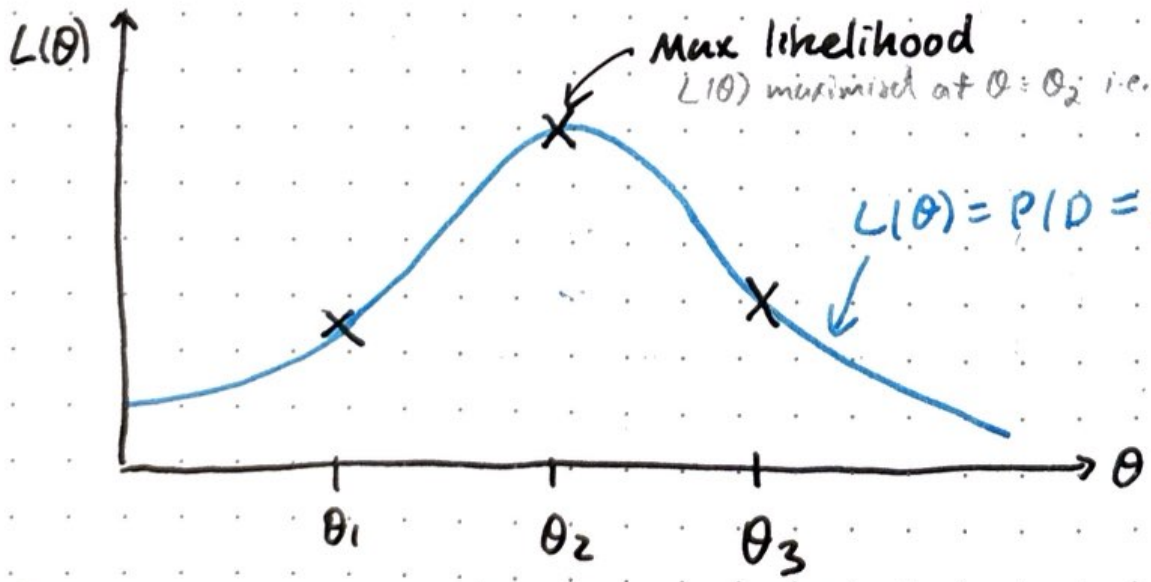


Sampling distribution

Likelihood: $P(D = D_{obs}|\theta) = L(\theta)$

$\int P(D|\theta)P(\theta)\,d\theta = P(D)$

N.B. $\int L(\theta)\,d\theta$ not necessarily $= 1$



Suppose we now observe data $\vec{D}_{obs}$ (outcome or realisation of $\vec{D}$), the likelihood is the sampling dist. evaluated at $\vec{D} = D_{obs}$, viewed as a fn. of the parameters i.e. $L(\vec{\theta}) = P(\vec{D} = D_{obs}|\vec{\theta})$

Notation:

merge?

Often we will elide the distinction between $\vec{D}$ and $\vec{D}_{obs}$

$P(\vec{D}|\vec{\theta})$ understood to mean $P(\vec{D}=\vec{D}_{obs}|\vec{\theta})$.

log likelihood $\ell(\theta) = \ln L(\theta)$.

iid case:

$$x \overset{iid}{\sim} f(x|\theta) \qquad i=1,\dots,N, \quad P(\vec{x}=(x_1,\dots,x_N)|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

individual likelihood

$$L(\theta) = P(\vec{x}|\vec{\theta}) = \prod_{i=1}^{N} f(x_i|\theta)$$

↑ full likelihood

ℓ

$$\ell(\theta) = \sum_{i=1}^{N} \ln f(x_i|\theta)$$ log likelihood

Fisher Information ⟷ Uncertainty

side note: does mean likelihood estimate score = 0 generally or not in n?)

$$L(\theta) = P(x|\theta)$$

DEFINE:

$$\text{Score} = S = \frac{\partial}{\partial\theta} \ln L(\theta) = \frac{\partial}{\partial\theta} \ln P(x|\theta)$$ defining S

↗ wrt $P(x|\theta)$

$$\mathbb{E}[S] = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln P(x|\theta)\right] = \int \left[\frac{\partial}{\partial\theta} \ln P(x|\theta)\right] P(x|\theta)\, dx$$

expectation over sample space

$$= \int \left[\frac{1}{P(x|\theta)} \frac{\partial}{\partial\theta} P(x|\theta)\right] P(x|\theta)\, dx$$

$$\overset{*}{=} \frac{\partial}{\partial\theta} \int P(x|\theta)\, dx = \frac{\partial}{\partial\theta}[1] = 0$$

* Under regularity conditions $\frac{\partial}{\partial\theta} \longleftrightarrow \int dx$

Fisher Information = Variance of the Score

$(I(\theta) = \mathrm{Var}(S) = E(S^2) - E(S)^2)$ → =0 from previous results

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln P(x|\theta)\right)^2\right] - \left(E\left[\frac{\partial}{\partial\theta}\ln P(x|\theta)\right]\right)^2$$

$$\stackrel{?}{=} -E\left[\frac{\partial^2}{\partial\theta^2}\ln P(x|\theta)\right]$$

$$E\left[\frac{\partial^2}{\partial\theta^2}\ln P(x|\theta)\right] = E\left[\frac{\partial}{\partial\theta}\left[\frac{1}{P(x|\theta)}\frac{\partial}{\partial\theta}P(x|\theta)\right]\right]$$

$$= E\left[-\frac{1}{P(x|\theta)^2}\frac{\partial}{\partial\theta}P(x|\theta)\frac{\partial}{\partial\theta}P(x|\theta) + \frac{1}{P(x|\theta)}\frac{\partial^2}{\partial\theta^2}P(x|\theta)\right]$$

$$= E\left[-\left(\frac{\partial}{\partial\theta}\ln P(x|\theta)\right)^2 + \frac{1}{P(x|\theta)}\frac{\partial^2}{\partial\theta^2}P(x|\theta)\right]$$

$$\int\left[\frac{1}{P(x|\theta)}\frac{\partial^2}{\partial\theta^2}P(x|\theta)\right]P(x|\theta)\,dx = \frac{\partial^2}{\partial\theta^2}(1) = 0$$

$$= -E\left[\left(\frac{\partial}{\partial\theta}\ln P(x|\theta)\right)^2\right]$$

$$= -I(\theta) \quad //$$

$$\boxed{I(\theta) = E\left[\frac{\partial^2}{\partial\theta^2}(-\ln P(x|\theta))\right]}$$

from wiki:
- amount of information R.V. X carries about an unknown parameter θ (upon which the probability of X depends)

$I(\theta) \geq 0$

Formally:
- variance of score
or - expected value of observed information

F.I.
→ used to calculate covariance matrices associated with MLE

bound on variance of estimators

Cramer - Rao Lower Bound

wiki: "inverse of fisher info. is a lower bound on the variance of any unbiased estimator"

Suppose we have estimator $T(x)$ for θ,

$$\mathbb{E}[T] = \theta + b(\theta).$$

Consider $\mathrm{Cov}(S(x), T(x)) = \mathbb{E}[S(x)T(x)] - \mathbb{E}[S]\mathbb{E}[T]$ (score; $\mathbb{E}[S] = 0$)

$$= \mathbb{E}\left[T(x) \frac{1}{P(x|\theta)} \frac{\partial}{\partial\theta} P(x|\theta)\right]$$

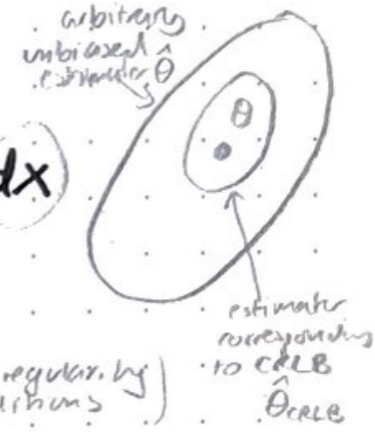
$$= \int \left[T(x) \frac{1}{P(x|\theta)} \frac{\partial}{\partial\theta} P(x|\theta)\right] P(x|\theta)\, dx$$

$$= \int \frac{\partial}{\partial\theta} T(x) P(x|\theta)\, dx$$

$$\overset{*}{=} \frac{\partial}{\partial\theta} \int T(x) P(x|\theta)\, dx$$

(* under regularity conditions)

$$= \frac{\partial}{\partial\theta} \mathbb{E}[T(x)] = 1 + b'(\theta)$$



$-1 \leq \mathrm{corr} \leq 1$

$|\mathrm{corr}| \leq 1$

$$\mathrm{Cov}(S,T) = \sqrt{\mathrm{Var}(S)}\sqrt{\mathrm{Var}(T)}\,\mathrm{Corr}(S,T)$$

$$|\mathrm{Cov}(S,T)| \leq \sqrt{\mathrm{Var}(S)}\sqrt{\mathrm{Var}(T)}$$

$$(1 + b'(\theta))^2 \leq \mathrm{Var}(S)\mathrm{Var}(T)$$

$$\leq I(\theta)\mathrm{Var}(T)$$

$$\boxed{\mathrm{Var}(T) \geq \frac{(1 + b'(\theta))^2}{I(\theta)}}$$

(wiki) why relevant to MLE?
→ MLE efficient → achieves CRLB as $N \to \infty$
→ ∴ no consistent estimator is more efficient than MLE
(asymptotically?)

CRLB gives lower bound on minimum possible variance for an estimator.
If there are 2 or more unbiased estimators, the one with the lowest variance is often preferred.

T unbiased → $b = 0$

$$\boxed{\mathrm{Var}(T) \geq I(\theta)^{-1}}$$

useful property

CRLB Multivariate Case

$\vec{x} \sim P(\vec{x}|\vec{\theta})$

and $\vec{T}(\vec{x})$ is unbiased estimator for $\vec{\theta}$,

$E[\vec{T}(\vec{x})] = \vec{\theta}$, $E[T_j(\vec{x})] = \theta_j$

Then $\left[\underset{\sim}{\text{Cov}}(\vec{T}(\vec{x})) \geq \underset{\sim}{I}^{-1}(\theta) \right]$

has elements $\text{Cov}(T_j, T_k)$

Fisher Matrix $\underset{\sim}{I}(\theta)$ has elements

$$I_{jk} = E\left[-\underbrace{\frac{\partial^2}{\partial\theta_j \partial_k} \ln P(\vec{x}|\vec{\theta})}_{\text{Hessian}} \right]$$

In particular:

$$\left[\text{Var}(T_i) \geq [I^{-1}(\theta)]_{ii} \right]$$

(diagonal terms)

$$H(\theta) = \begin{vmatrix} \frac{\partial^2 \ell}{\partial\theta_1^2} & \frac{\partial^2 \ell}{\partial\theta_1 \partial\theta_2} & \cdots \\ \frac{\partial^2 \ell}{\partial\theta_2 \partial\theta_1} & \ddots & \\ \vdots & & \end{vmatrix}$$

(all eval at $\theta = \hat{\theta}$)

Max Likelihood

$$X_i \overset{iid}{\sim} f(x|\theta) \quad i = 1, \ldots, N$$

$$L(\theta) = P(\vec{x}|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\, L(\theta)$$

$$= \underset{\theta}{\text{argmax}}\, \ln L(\theta)$$

(i think?) to find $\hat{\theta}_{MLE}$:

$\frac{\partial}{\partial\theta} \ell(\theta) = 0$

multiparameter do $\frac{\partial}{\partial\theta_1} \ell(\theta_1) = 0$

$\frac{\partial}{\partial\theta_2} \ell(\theta_2) = 0$

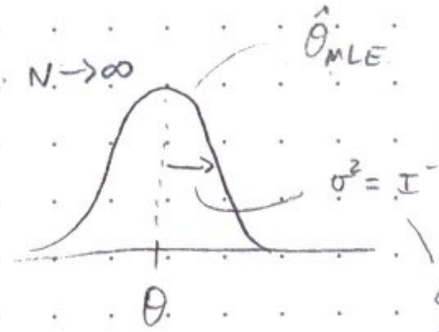
etc.

MLE Properties (Model is true!)

* Consistent (strongly?): $\hat{\theta}_{MLE} \to \theta_{true}$ as $N \to \infty$
* Asymptotically unbiased: $E[\hat{\theta}_{MLE}] \to \theta_{true}$ as $N \to \infty$
* Asymptotically Normal $(\hat{\theta}_{MLE} - \theta_{true}) \xrightarrow{d} N(\theta, I^{-1})$?

↳ Asymptotic normality: consistent estimators $\hat{\theta}$ have dist around true θ that approaches a normal dist. (variance decreasing as $1/N$)

↳ why gaussian is useful: gaussian covariance matrices associated w/ MLE



achieves CRLB as $N \to \infty$ (efficient!)

no other estimator is more efficient

Lecture 7 7.2.25

Max Likelihood Properties (cont.)

$$X_i \overset{iid}{\sim} f(x|\theta) \qquad i = 1, \ldots, N$$

$$L(\theta) = P(\vec{x}|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

$$MLE \rightarrow \hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\, L(\theta) = \underset{\theta}{\text{argmax}}\, \ln L(\theta)$$

* **Consistent:** $\hat{\theta}_{MLE} \rightarrow \theta_{true}$ as $N \rightarrow \infty$

$(\forall \varepsilon > 0,\ Pr(|\hat{\theta}_{MLE} - \theta_{true}| > \varepsilon) \rightarrow 0 \text{ as } N \rightarrow \infty)$

* **Asymptotically Unbiased:** $E[\hat{\theta}_{MLE}] \rightarrow \theta_{true}$ as $N \rightarrow \infty$ (but not necessarily unbiased)

* **Asymptotically Normal:** $(\hat{\theta}_{MLE} - \theta_{true}) \xrightarrow{d} N(0, I^{-1})$

$$I(\theta) = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right] = -N E\left[\frac{\partial^2 \ln f}{\partial \theta^2}\right]$$

↖ Expected Fisher Information

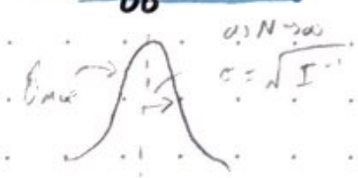
$$\hat{I} = \text{"observed Fisher Info."} = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta = \hat{\theta}_{MLE}}$$

$\hat{I} \rightarrow I(\theta)$ as $N \rightarrow \infty$ (⇒ asymptotically normally)

* **Efficient:** Asymptotically achieves CRLB

$\text{Var}(\hat{\theta}_{MLE}) \rightarrow I^{-1}$ as $N \rightarrow \infty$

"saturates"



also called "equivariance property"

* **Functionally invariant:** $\alpha = g(\theta)$ ← (parameter transf. g)

$\hat{\alpha}_{MLE} = g(\hat{\theta}_{MLE})$

If $\hat{\theta}_{MLE}$ is MLE for θ and $g(\theta)$ is bijective transf. of θ then MLE for $\alpha = g(\theta)$ is $\hat{\alpha}_{MLE} = g(\hat{\theta}_{MLE})$

Multiparameter Case MLE

$$L(\vec{\theta}) = p(\vec{x}|\vec{\theta}) = \prod_{i=1}^{N} p(x_i|\vec{\theta}) \qquad \vec{\theta}_{MLE} = \underset{\vec{\theta}}{\operatorname{argmax}}\, L(\vec{\theta})$$

* **Asymptotically Normal:** $\vec{\theta}_{MLE} - \vec{\theta}_{true} \xrightarrow{d} N(\vec{0}, \underset{\sim}{I}^{-1})$

↑ Inverse Fisher Matrix

I_{jk} = Exp. Fisher Info Matrix

$$= \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k}\right] = -N\,\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial\theta_j \partial\theta_k}\right]$$

hessian?

similarly can define

Observed F.I. $= \hat{I}_{jk} = -\left.\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k}\right|_{\vec{\theta}=\vec{\theta}_{MLE}}$

$\approx I_{jk}$ as $N \to \infty$

* **Asymptotically Efficient:** $\mathrm{Cov}(\vec{\hat{\theta}}_{MLE}) \to \underset{\sim}{I}^{-1}$

In particular,

usually we want to know variance of particular components →

$\mathrm{Var}(\theta_{MLE,i}) \to (I^{-1})_{ii}$ (diagonal)

Evaluate Fisher Matrix

Quick Example:

$X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \dots, N$

$\vec{\theta} = (\mu, \sigma^2)$

$$L(\vec{\theta}) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2/\sigma^2}$$

$$\rightarrow \ell(\hat{\theta}) = \sum_{i=1}^{N} -\tfrac{1}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}$$

(1st deriv)

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\rightarrow \hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} \rightarrow Var(\hat{\mu}_{MLE}) = \sigma^2/N$$

(2nd deriv)

$$\frac{\partial^2 \ell}{\partial \mu^2} = \sum_{i=1}^{N} -1/\sigma^2 = -N/\sigma^2$$

(1st deriv wrt σ^2)

$$\frac{\partial \ell}{\partial \sigma^2} = \sum_{i=1}^{N} -\tfrac{1}{2}\frac{1}{\sigma^2} + \tfrac{1}{2}\frac{(x_i-\mu)^2}{(\sigma^2)^2} = 0$$

$$\rightarrow \hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i-\bar{x})^2$$

(asymptotically unbiased but not unbiased)

(N.B. this is biased! $\mathbb{E}[\hat{\sigma}^2_{MLE}] = \frac{N-1}{N}\sigma^2$) — see wiki, 'variance' (1) biased sample variance

But sample variance $S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$

is unbiased for σ^2.

Remember! calculus biases/expectations remember multiply out $(x-\bar{x})^2$ first & remember sums to consider dep/indep variables i.e. split $(x_1+x_2+\dots)^2 = x_1^2 + x_1x_2 + x_1x_3$ ... (not all zero!)

(2nd deriv wrt σ^2)

$$\frac{\partial^2 \ell}{\partial^2 \sigma^2} = \sum_{i=1}^{N} \frac{1}{2\sigma^4} - \frac{(x_i-\mu)^2}{\sigma^6}$$

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2}\right] = -\frac{N}{2\sigma^4} \quad \left(\text{using } \mathbb{E}[(x_i-\mu)^2] = \sigma^2\right)$$

can check yourself

Cross term $\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2 \partial \mu}\right] = 0$ (using $\mathbb{E}[x_i - \mu] = 0$)

(actually $\mathbb{E}\{-\frac{\partial^2 \ell}{\partial \mu^2}\}$ etc?).

$$\underset{\sim}{I} = \begin{pmatrix} -\frac{\partial^2 \ell}{\partial \mu^2} & -\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} \\ -\frac{\partial^2 \ell}{\partial \sigma^2 \partial \mu} & -\frac{\partial^2 \ell}{\partial^2 \sigma^2} \end{pmatrix} = \begin{pmatrix} N/\sigma^2 & 0 \\ 0 & N/2\sigma^4 \end{pmatrix}$$

$$\longrightarrow \quad \underset{\sim}{I}^{-1} = \begin{pmatrix} \sigma^2/N & 0 \\ 0 & 2\sigma^4/N \end{pmatrix}$$

$\hat{\mu}$

$$\mathrm{Var}(\hat{\mu}_{MLE}) = \sigma^2/N \overset{?}{\geq} (I^{-1})_{11}$$

$$\mathrm{Var}(\hat{\sigma}^2_{MLE}) \overset{?}{\geq} (I^{-1})_{22}$$

check this yourself
is variance greater than our bound from MLE theory?

$\mathrm{Var}(\hat{\mu}_{MLE}) = (I^{-1})_{11}$ as $N \to \infty$ ✓ as expected

Laplace Approximation:

recall $\ell(\theta) = \ln L(\theta)$

score = 0 because we maximize to find MLE

$-\hat{I}$

taylor approx around MLE:

$$\ell(\theta) \simeq \ell(\hat{\theta}_{MLE}) + \cancel{\frac{\partial \ell}{\partial \theta}\Big|_{\hat{\theta}_{MLE}}}^{0} (\theta - \hat{\theta}_{MLE}) + \frac{1}{2}\frac{\partial^2 \ell}{\partial \theta^2}\Big|_{\hat{\theta}_{MLE}} (\theta - \hat{\theta}_{MLE})^2 + \dots$$

Asymptotically, $N \to$ large

$$\ell(\theta) \simeq \ell(\hat{\theta}_{MLE}) - \frac{1}{2}\hat{I}(\theta - \hat{\theta}_{MLE})^2$$

$$L(\theta) \simeq L(\hat{\theta}_{MLE})\, e^{-\frac{1}{2}(\theta - \hat{\theta}_{MLE})^2 / \hat{I}^{-1}}$$

(Gaussian shape)

Multiparameter case

$$L(\vec{\theta}) \simeq L(\hat{\vec{\theta}}_{MLE}) \times e^{-\frac{1}{2}(\vec{\theta} - \hat{\vec{\theta}}_{MLE})^T \underline{\underline{\hat{I}}}^{-1} (\vec{\theta} - \hat{\vec{\theta}}_{MLE})}$$

Reminder:

multidim taylor expansion

$$f(\underline{x}) \simeq f(\underline{a}) + \underline{\nabla} f(\underline{a}) \cdot (\underline{x} - \underline{a}) + \frac{1}{2}(\underline{x} - \underline{a})^T \underline{\underline{H}} f(\underline{a}) (\underline{x} - \underline{a}) + \dots$$

grad f at a

Hessian of f at a

e.g. $f(x,y) \simeq f(a,b) + \frac{\partial f}{\partial x}\Big|_{a,b}(x-a) + \frac{\partial f}{\partial y}\Big|_{a,b}(y-b) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\Big|_{a,b}(x-a)^2 + \frac{\partial^2 f}{\partial x \partial y}\Big|_{a,b}(x-a)(y-b) + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}\Big|_{a,b}(y-b)^2 + \dots$

Example (slides): Calibrating type Ia supernova absolute magnitudes with intrinsic dispersion and heteroskedastic measurement error. (useful for ex sheet 1)

Determining Astronomical Distances using Standard Candles → key step in determining hubble constant

1. Estimate or model luminosity L of a class of astronomical objects
2. Measure apparent brightness or flux F
3. Derive distance D to object using inverse square law

$$F = L / 4\pi D^2$$

4. Optical astronomer's units $\mu = m - M$

m = apparent magnitude (log apparent brightness, flux)
M = absolute magnitude (log luminosity)
μ = distance modulus (log distance)

can use hubble to find distances

Hubble: $V = H_0 \times$ distance
distance $\propto$ velocity (redshift)

Hubbles law
$z \approx v/c$

Now use type Ia supernovae as standard candles

We want to measure the hubble constant
→ can use (local) distance ladder → hubble tension is discrepancy between using this method & another one!)

2nd tier of distance ladder:
Can use cepheid stars: period of rotation linked to luminosity (pulsating)

1st tier: geometry, 2nd tier: cepheids, 3rd tier: supernovae
→ use inbetween tiers to calibrate e.g. same distance cepheids & SNIa
cf. 3 rungs of the distance ladder

calibrate when they are in the same galaxy

hubble flow = things that are further away tend to move away from us faster

Lecture 8

10.2.25

Recap example (slides)

- Measure apparent brightness or flux F
- Derive distance to object using inverse square law

$$F = L/4\pi D^2$$

- key step in determining hubble constant ($\ell = \frac{\text{redshift}}{D}$)
- 3 rungs to distance ladder: geometry, cepheids, supernovae (can calibrate between rungs)

we will focus on cepheids

back on board!

Calibrating SNIa Abs Mags (board)

Calibrator sample we observe: N SNe, $s = 1, \dots, N$

m_s = true apparent magnitude

$\hat{m}_s$ = measured apparent magnitude

$$\hat{m}_s \sim N(m_s, \sigma^2_{m,s})$$ ← given in data table

equivalently, $\hat{m}_s = m_s + \varepsilon_{ms}$ ← $\sim N(0, \sigma^2_{ms})$ ← data

Cepheid distance estimates to SN galaxy:

μ_s = true distance modulus

$$= 25 + 5\log_{10}\left(\frac{\text{distance}}{\text{Mpc}}\right)$$

again measurement error model → $\hat{\mu}_s \sim N(\mu_s, \sigma^2_{\mu,s})$ ← given in data table

equiv. $\hat{\mu}_s = \mu_s + \varepsilon_{\mu s}$ ← $\sim N(0, \sigma^2_{\mu s})$

↑ "data"

Abs Mag Distribution (Population):

θ → what we want to estimate

$$M_s \sim N(M_0, \sigma_{int}^2)$$

(σ_{int}^2 unknown)

intrinsic error (not like measurement error) intrinsic due to different masses & gas, properties etc

equiv. $M_s = M_0 + \varepsilon_{int,s} \rightarrow \sim N(0, \sigma_{int}^2)$

like before, want to estimate θ, so want to form likelihood eq:

$M_s = M_s + \mu_s$ (Latent variable equation) (i.e variables without hats relate the latent variables)

Define: $\hat{M}_s = \hat{m}_s - \hat{\mu}_s$ (Estimated Abs Mag)

$= m_s + \varepsilon_{m,s} - \mu_s - \varepsilon_{\mu,s}$

$= M_s + \varepsilon_{m,s} - \varepsilon_{\mu,s}$

$= M_s + \varepsilon_{err,s} \sim N(0, \underbrace{\sigma_{m,s}^2 + \sigma_{\mu,s}^2}_{\sigma_{err,s}^2})$

(combine errors) use: variance of sum of gaussian RVs is sum of variance

also can define $\hat{M}_s = M_0 + \varepsilon_{int,s} + \varepsilon_{err,s}$ (mean)

$\hat{M}_s \sim N(M_0, \sigma_{int}^2 + \sigma_{err,s}^2)$

b/c ε have 0 mean

combine variances (M_0 0 variance)

Likelihood:

(note $\hat{M}_s$ not i.i.d (but independent) aka heteroskedastic)

(diff SN in diff galaxies so indep but not i.i.d. identically dist)

given indep. between SNe (independence)

$$P(\hat{M}_s | M_0, \sigma_{int}^2) = N(\hat{M}_s | M_0, \sigma_{int}^2 + \sigma_{err,s}^2)$$

heteroskedastic measurement error

atypical in cosmology

$$L(\theta) = P(\hat{M} | M_0, \sigma_{int}^2) = \prod_{s=1}^{N} N(\hat{M}_s | M_0, \sigma_{int}^2 + \sigma_{err,s}^2)$$

example of heteroskedastic error & multiple sources of uncertainty

$\sigma_{err,s}$ creates problem for MLE, cannot be solved analytically

so solve numerically

(see slides for example code)

cannot solve MLE analytically!

Supernova Absolute Magnitude Distribution: Selection Effects (board)

same ex. but now

Assume measurement error $= 0 = \sigma_{m,s} = \sigma_{\mu,s}$ (for simplicity & argument)

Population variability $\sigma_{int} \to \sigma$.

Only intrinsic population is source of variability

$$(\sigma^2 = \sigma_{int}^2).$$

Suppose distance is the same for an entire sample of SN

$$\mu_s = \mu .$$

population dist

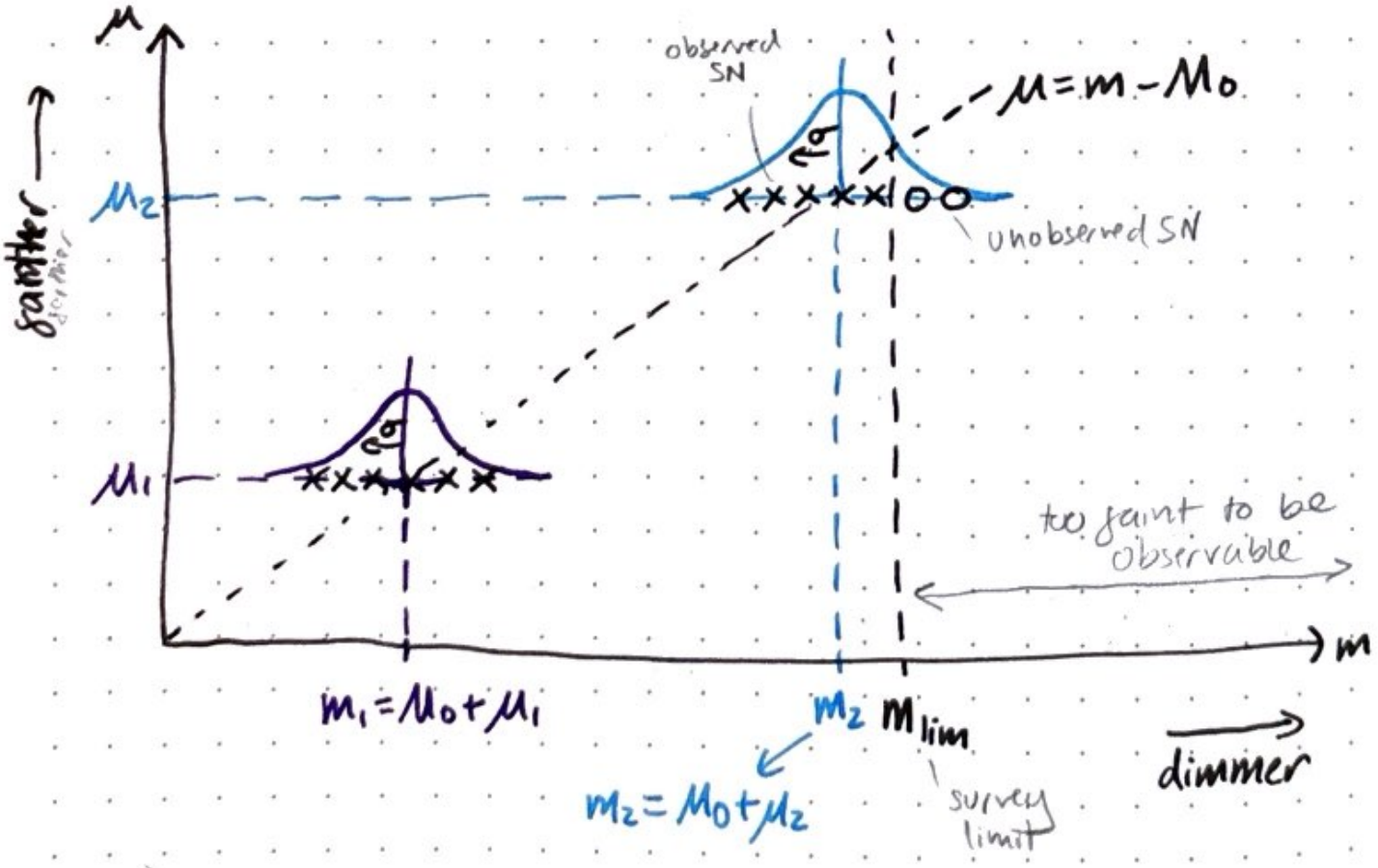
$$M_s \overset{iid}{\sim} N(M_0, \sigma^2)$$

latent variable eq

$$(m_s = M_s + \mu_s)$$

combine

$$\to m_s \overset{iid}{\sim} N(M_0 + \mu, \sigma^2), \quad s = 1, \ldots, N$$



Volz stops)

survey can only detect up to certain magnitude limit, can't see anything dimmer than this

Suppose we have 2 populations SNe: M_1 & M_2. Assign magnitudes based similar but in the 2nd we don't see SNe above some limit, what do we do? →

Data: $\{m_s\} = \vec{m}$

fixed μ for sample

Naive Likelihood: $P(m_s | M_0, \sigma^2) = N(m_s | \mu + M_0, \sigma^2)$

$$L(M_0, \sigma^2) = \prod_{s=1}^{N} P(m_s | M_0, \sigma^2)$$

↓ naively apply naive likelihood

If $M_0 + \mu$ close to m_{lim}, MLE will be biased:

- $\hat{M}_0$ too bright — (b/c didn't see the dim supernovae)
- $\hat{\sigma}$ too small — (b/c tail of dist is cut off, apparent width smaller than true width)

how do we solve this?

Accounting For Selection Effects (sheet 1 q2)

Let I_s (indicator variable) $= \begin{cases} 1, & \text{if SN observed} \\ 0, & \text{if SN NOT observed} \end{cases}$

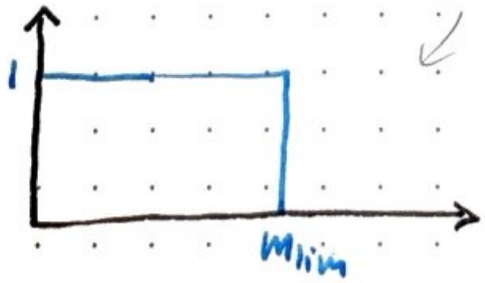
Define Selection Function:

prob indicator given data

$$P(I_s | m_s) = \begin{cases} 1, & m_s \leq m_{lim} \\ 0, & m_s \geq m_{lim} \end{cases}$$

$$= 1 - H(m_s - m_{lim})$$

↙ heaviside step fn



Lecture 9

12.2.25

(brief recap of SN problem setup on board - useful for sheet 1) example!

SELECTION EFFECTS cont.

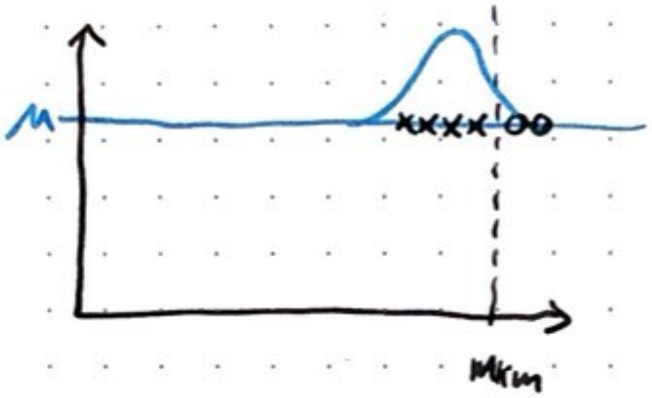
recap:

$M_s = N(M_0, \sigma^2)$

$s = 1, \dots, N$

$\mu_s = \mu$ (known)

known precisely so no measurement error



Formulate likelihood to account for selection effects:

Let $I_s = \begin{cases} 1, & \text{if SN observed} \\ 0, & \text{if SN NOT observed} \end{cases}$

Selection function:

$P(I_s = 1 | m_s)$:

$$S(m_s) = P(I_s | m_s) = \begin{cases} 1, & m_s < m_{lim} \\ 0, & m_s \geq m_{lim} \end{cases}$$

$1 - H(m_s - m_{lim})$

Observed Data Likelihood:

$$P(m_s \mid I_s = 1, \theta) = \frac{P(I_s = 1, m_s \mid \theta)}{P(I_s = 1 \mid \theta)}$$

$$= \frac{P(I_s = 1 \mid m_s, \theta)\, P(m_s \mid \theta)}{\int P(I_s = 1 \mid m_s, \theta)\, P(m_s \mid \theta)\, dm_s}$$

$$= \frac{S(m_s)\, N(m_s \mid M_0 + \mu, \sigma^2)}{\int S(m_s)\, N(m_s \mid M_0 + \mu, \sigma^2)\, dm_s}$$

cut off above m_{lim}

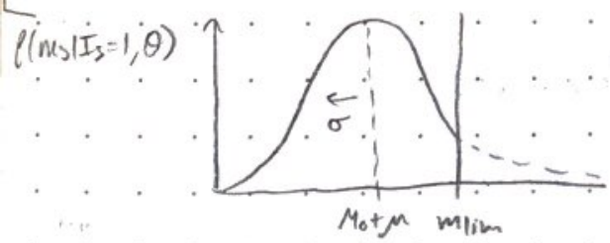
$$= \frac{[1 - H(m_s - m_{lim})]\, N(m_s \mid M_0 + \mu, \sigma^2)}{\int_{-\infty}^{m_{lim}} N(m_s \mid M_0 + \mu, \sigma^2)\, dm_s}$$

Gaussian CDF Φ ← denominator

(TRUNCATED NORMAL) "look this up on wikipedia"

$$= TN(m_s \mid M_0 + \mu, \sigma^2, -\infty, m_{lim})$$

- untruncated mean & variance
- lower truncation limit
- upper truncation limit

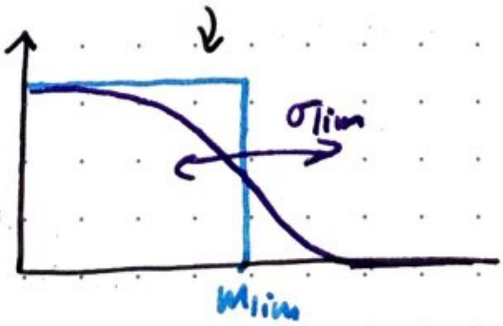


Challenge: what is $S(m_s) = P(I_s \mid m_s) = \Phi\left(\frac{m_{lim} - m_s}{\sigma_{lim}}\right)$?

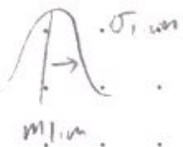
$P(I_s = 1 \mid m_s)$?

(more realistic selection pdf to take into account survey limit is not the same every night)

- limit not a sharp boundary. to take into account e.g. obscuring mist / seeing conditions clearer days



?



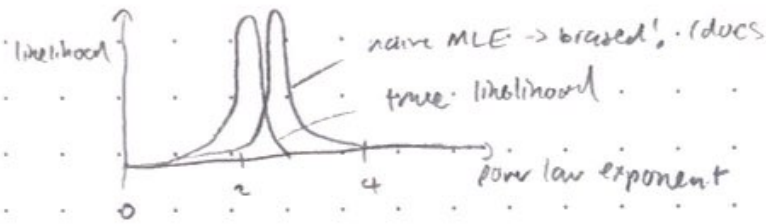
(slides)

Ex sheet 1 q 2: star formation in Perseus (clouds of dust collapse under grav. & form stars here)

→ in densest cloud regions stars are hidden

→ densest regions will be where largest stars form → selection effect

→ pareto or power law distribution $P(m) \propto m^{-\gamma}$ $(m \geq b_0)$



QUANTIFYING UNCERTAINTY USING BOOTSTRAP

Frequentist interpretation:

Consider variability of your estimator $g(\vec{x})$ for θ under (imaginary) repetitions of your experiment. (Random realisations of the potential data).

How does $g(\vec{x})$ behave under the potential datasets you did not observe?

under $x \sim P(x|\theta)$

$$\text{e.g. } \mathrm{Var}[g(\vec{x})] = \mathbb{E}[g(\vec{x}) - \mathbb{E}[g(\vec{x})]].$$

If $g(\vec{x})$ is approximately Gaussian distributed

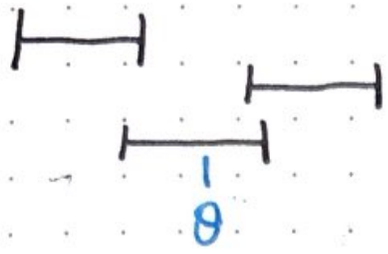
→ 68% confidence interval $g(\vec{x}) \pm \sqrt{\mathrm{Var}(g(\vec{x}))}$ (standard notation / shorthand)

$$\left[g(\vec{x}) - \sqrt{\mathrm{Var}(g(\vec{x}))},\ g(\vec{x}) + \sqrt{\mathrm{Var}(g(\vec{x}))}\right]$$

$(1-\alpha)$% confidence interval $[L(\vec{x}), U(\vec{x})]$ contains the true value θ_{true} in at least $(1-\alpha)$% of the realisations.

(↳ repeat experiments, fraction of intervals that contain true parameter will approach 68%)

NOT prob a particular interval contains the true parameter (common confusion)



get diff interval for each experiment ↑
68% confidence interval → 68% of these realizations contain θ

$[L(x), U(x)]$ is a random interval vs. $[L(x_{obs}), U(x_{obs})]$ evaluated on observed numerical values ↑, only one dataset x_{obs}! (either contains θ_{true} or doesn't)

Bootstrap:

Use the observed dataset to simulate the variability of the unobserved (imaginary) data sets.

BOOTSTRAP SAMPLE = sample with replacement from the observed dataset to the sample size

E.g. $x_1, \ldots, x_5 \overset{iid}{\sim} \text{Poisson}(\lambda)$

could do w/ max. likelihood but suppose i don't know exist

$$\rightarrow P(x_i) = \frac{\lambda^{x_i} e^{\lambda}}{x_i!}$$

Real data (observed) $\vec{x}_{obs} = (3, 8, 2, 4, 5)$

Suppose you want to estimate the skewness of $P(x)$ (asymmetry)

i.e. skewness = $\dfrac{\mathbb{E}[(x-\mu)^3]}{(\sigma^2)^{3/2}}$ (true mean, true variance) $= \lambda$

Sample skewness $g(\vec{x}) = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)^3}$

Bootstrap B "replicate" datasets from observed dataset:

"sample w/ replacement" from original e.g. sample N times from 3,8,2,4,5 gives 2,5,4,4,4

$\vec{x}^{obs} = (3,8,2,4,5)$ $\hat{g}^{obs} = g(\vec{x}^{obs}) = 0.6927$

$\vec{x}^{b=1} = (2,5,4,4,4)$ $\hat{g}_1 = g(\vec{x}^{b=1}) = -0.8625$

$\vec{x}^{b=2} = (2,4,2,8,8)$ $\hat{g}_2 = g(\vec{x}^{b=2}) = 0.2115$

$\vec{x}^{b=3} = (5,2,8,2,5)$ $\hat{g}_3 = g(\vec{x}^{b=3}) = 0.3436$

$\vdots$

$\vec{x}^{b=B} =$ $\hat{g}_B =$

Can now compute sample variance

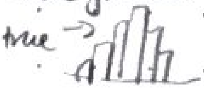
$\hat{Var}(\{\vec{g}_1, \dots, \vec{g}_B\})$ $\rightarrow \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$ (?)

Standard error $= \sqrt{\hat{Var}}$

$\hat{g} = 0.6927 \pm 0.635 \approx 68\%$ C.I.

(slides)
back to stergamatn pesevs example:
how do i get standard error on exponent?
→ compare fisher matrix approach and compare this against bootstraps
- fit model to each bootstrapped realisation

plot max likelihood estimates for each as histograms



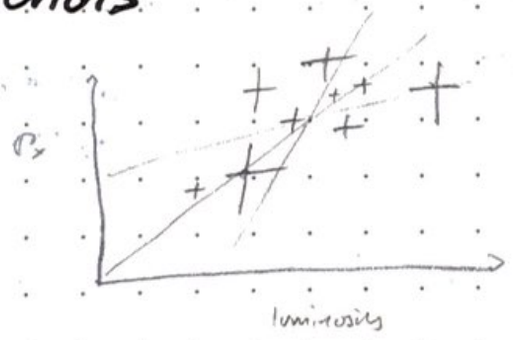
(false confidence → smaller uncertainty than true)

Lecture 10

14.2.25

REGRESSION (slides)

* Fit a function $E[y|x] = f(x;\theta)$ for the mean relation between y and x
* Basic approaches
  - → ordinary least squares (homoskedastic scatter)
  - → generalised least squares (heteroskedastic, correlated scatter)
  - → weighted least squares (minimum χ^2, known variance)
  - → maximum likelihood
* Real data problems require more complex modelling
  - → regression dilution from covariate measurement errors



e.g.
each quasar has diff measurement error
regression distinguishably apply OLS gives biased slope when there are errors in x (sheet 2)

Ordinary least squares (OLS):

Linear model $$y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{ij} + \varepsilon_i$$

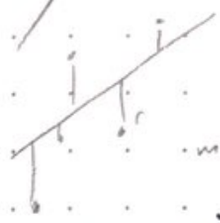
$$\downarrow$$

$$Y = X\beta + \varepsilon$$

$i = 1, \ldots, N$ objects, $E[\varepsilon_i] = 0$, homoskedastic
$\mathrm{Var}[\varepsilon_i] = \sigma^2$ (known).

(don't think this is too important to memorise? (slides))

$$y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{ij} + \varepsilon_i$$



assume no error in y, notice error in x

minimises sum squared of these distances

residual sum of squares

Minimise w.r.t. β (solve for gradient = 0):

$$RSS = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j}^{k-1} \beta_j x_{ij})^2 = (Y - X\beta)^T (Y - X\beta)$$

("simple linear regression" era wskei) for ez example

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$ just accept this?

$$Var(\hat{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1}$$

unbiased & has minimum variance under these assumptions

$$\mathbb{E}(\hat{\beta}_{OLS}) = \beta \text{ (unbiased)}$$

BLUE

best linear unbiased estimate

Estimate unknown variance ← OLS estimate of variance is $\hat{\sigma}^2 = \frac{RSS}{N-2}$ (for 2 param example)

$$\hat{\sigma}^2 = \frac{1}{N-k} (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

$k=2$ for simple linear $y = \alpha + \beta x$? (+ε)

$-k$ account for fact we have already estimated $\hat{\beta}$'s from data (k coefficients)

(WEIGHTED LEAST SQUARES) – aka χ^2 minimisation:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \beta_0 - \sum_{j}^{k-1} \beta_j x_{ij})^2}{\sigma_i^2}$$ ← known!

χ^2 r.v. = sum of squared Gaussian r.v.s

Minimise w.r.t. β:

If Gaussian errors, at $\beta = \beta_{min}$, $\chi^2 \sim \chi^2_{N-k}$.

Model check: $\mathbb{E}(\chi^2_{N-k}) = N-k$ → (if assumptions true (contrary))

$$\frac{\chi^2}{N-k} \approx 1 \quad \text{(for large } N-k\text{)}$$

χ^2 statistic should have a χ^2 distribution with $N-k$ deg of freedom (if gaussian errors?)

reduced χ^2 statistic: a way to test model fit

These are special cases of generalised least squares.

Linear model $y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{ij} + \varepsilon_i$, $i = 1, \ldots, N$ objects

$\downarrow$

$Y = X\beta + \varepsilon$

$\mathbb{E}[\varepsilon_i] = 0$, correlated errors

$\text{Var}[\varepsilon] = \text{Cov}[\varepsilon, \varepsilon^T] = W$ (known)

Minimise w.r.t. β

$RSS = (Y - X\beta)^T W^{-1} (Y - X\beta)$

$\hat{\beta}_{GLS} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$

$\mathbb{E}[\hat{\beta}_{GLS}] = \beta$ (unbiased)

$\text{Var}[\hat{\beta}_{GLS}] = (X^T W X)^{-1}$

They can also be thought of as maximum likelihood (assuming Gaussian errors)

$Y = X\beta + \varepsilon$, $Y \sim N(X\beta, W)$

Maximise w.r.t. β (solve for gradient = 0)

$L(\beta) = P(Y|\beta, X) = N(Y|X\beta, W)$

do more likelihood on β (and also W)

(end of slides)

(minimize χ^2 estimation: grid values of θ that make $\chi^2(\theta)$ as small as possible)

χ^2 MINIMISATION (boord) (vs max likelihood)

"weighted least squares": errors not iid (but still indep) -> weight according to true σ_i^2 so that line doesn't get dragged to points w/ large σ_i

(x_i, Y_i) , $i = 1, \ldots, N$

Case: Variance unknown

$$Y_i = g(x_i; \theta) + \varepsilon_i \quad , \quad \mathrm{Var}(\varepsilon_i) = \sigma_i^2 \text{ known}$$

$$= a + bx + \varepsilon_i \quad \text{(linear)}$$

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(Y_i - g(x_i;\theta))^2}{\sigma_i^2}$$

$$\hat{\theta} = \arg\min_{\theta} \chi^2$$

Relation to maximum likelihood:

$$Y_i = g(x_i;\theta) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_i^2)$$

$$P(Y_i | X_i) = N(Y_i | g(x_i;\theta), \sigma_i^2)$$

$$L(\theta) = \prod_{i=1}^{N} P(Y_i | X_i; \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}(Y_i - g(x_i;\theta))^2/\sigma_i^2}$$

$$-2 \ln L(\theta) = \sum_{i=1}^{N} \ln(2\pi\sigma_i^2) + \sum_{i=1}^{N} \frac{(Y_i - g(x_i;\theta))^2}{\sigma_i^2}$$

$$= \sum_{i=1}^{N} \ln(2\pi\sigma_i^2) + \chi^2(\theta) \quad = \chi^2(\theta) + \text{const.}$$

(min $\chi^2(\theta)$ same as max likelihood)

→ If σ_i known: $\hat{\theta}_{\min \chi^2} = \hat{\theta}_{MLE}$

→ If σ_i^2 (or a component of it) unknown then NOT true!

(note: go gaussian χ^2, ML is same as LS)

Case: Variance component unknown: (unknown)

$$Y_i = g(x_i;\theta) + \varepsilon_{int}^{\,i} + \varepsilon_m^{\,i} \quad \Big| \quad \varepsilon_{int}^{\,i} \sim N(0, \sigma_{int}^2), \quad \varepsilon_m^{\,i} \sim N(0, \sigma_{m,i}^2)$$

($\sigma_{m,i}$ known)

$$\chi^2(\theta, \sigma_{int}^2) = \sum_{i=1}^{N} \frac{(Y_i - g(x_i;\theta))^2}{\sigma_{int}^2 + \sigma_{m,i}^2}$$

Problem: χ^2 is minimised when $\sigma_{int} \to \infty$!!

Max likelihood:

$$P(Y_i | x_i, \theta, \sigma_{int}^2) = N(Y_i | g(x_i;\theta), \sigma_{m,i}^2 + \sigma_{int}^2)$$

$$L(\theta, \sigma_{int}^2) = \prod_{i=1}^{N} \left[2\pi(\sigma_{int,i}^2 + \sigma_{m,i}^2)\right]^{-1/2} e^{-1/2 \frac{(Y_i - g(x_i;\theta))^2}{\sigma_{int}^2 + \sigma_{m,i}^2}}$$

$$-2\ln L(\theta, \sigma_{int}^2) = \sum_{i=1}^{N} \frac{(Y_i - g(x_i;\theta))^2}{\sigma_{int}^2 + \sigma_{m,i}^2} + \ln\left[2\pi(\sigma_{int}^2 + \sigma_{m,i}^2)\right]$$

(1) (2)

When $\sigma_{int} \to$ large, (1) ↓ but (2) ↑.

Max likelihood → finite estimate of σ_{int}!

i.e. assuming gaussian, independent & known errors → convenient but not derivable from first principles. likelihood is more broadly applicable

Lesson: χ^2 is an ad-hoc prescription. Likelihood is derived from explicit modelling assumptions.

slides: OLS leads to regression dilution when there are errors in x (usually measurement errors) which causes slope to be shallower than true value aka regression dilution

OLS gives shallow slope if you have error on x data. (regression dilution)

Probabilistic Generative Modelling (slides)

$P(D|\alpha, \theta)$

* Forward model comprises series of probabilistic steps describing conceptually how the observed data was generated from the parameters of interest

* Can introduce intermediate parameters / unobserved latent variables α (e.g. true values corresponding to the observed data)

  "what data would be had there been no measurement error"

* From forward model, derive the sampling distribution e.g.

$$P(D|\theta) = \int P(D|\alpha) P(\alpha|\theta) \, d\alpha$$

integrate out to leave params of interest

* Using observed data D, draw inference from likelihood function

Dons

$$L(\theta) = P(D|\theta)$$

* Or if Bayesian with prior $P(\theta)$: sample posterior

$$P(\theta|D) \propto P(D|\theta) P(\theta)$$

params: θ want
α nuisance

$$\left[P(\theta|D) = \int P(\theta, \alpha|D) \, d\alpha \right]$$

ex ways:

$$\Rightarrow \quad P(D|\theta) = \int P(D|\theta, \alpha) P(\alpha|\theta) \, d\alpha \quad \checkmark$$

show as before since w/ $P(\theta|D) = \int P(\theta, \alpha|D) \, d\alpha$ ✓

(how do i come up with model for this complex situation where I have measurement error in my x and y data)

Generative Model (slides) (same example from L3)

- Population distribution $\xi \sim N(\mu, \tau^2)$
- Regression Model $\eta_i|\xi_i \sim N(\alpha + \beta\xi_i, \sigma^2)$
- Measurement Error $[x_i, y_i]|\xi_i, \eta_i \sim N([\xi_i, \eta_i], \Sigma)$

For this example let $\Sigma = \begin{pmatrix} \sigma_{x,i}^2 & 0 \\ 0 & \sigma_{y,i}^2 \end{pmatrix}$ (for simplicity)

$x_i = \xi_i + \varepsilon_{x,i}$ ← $\sigma_{x,i}^2$

$y_i = \eta_i + \varepsilon_{y,i}$ ← $\sigma_{y,i}^2$

probabilistic generative model for linear regression

- Population dist. indep. variable $\psi = (\mu, \tau)$
- Regression Parameters $\theta = (\alpha, \beta, \sigma^2)$
- Latent (true) variables (i.e. no measurement error) (ξ_i, η_i)
- Observed data with measurement uncertainties $(\sigma_{x,i}, \sigma_{y,i})$ → $\Sigma = \begin{pmatrix} \sigma_{x,i}^2 & 0 \\ 0 & \sigma_{y,i}^2 \end{pmatrix}$

now can

Generate observed data from latent variables.

(see slides -> same as L3)

(now i've introduced all these new parameters how do i get rid of them?)

derive this on sheet 2a (same as slides) general solution

Formulating likelihood Function: Marginalising latent variables

$$P(x_i, y_i | \theta, \psi) = \iint P(x_i, y_i, \xi_i, \eta_i | \theta, \psi) \, d\xi_i \, d\eta$$

"observed data likelihood" — "complete data likelihood"

$$P(x_i, y_i | \theta, \psi) = \iint P(x_i, y_i | \xi_i, \eta_i) \, P(\eta_i | \xi_i, \theta) \, P(\xi_i | \psi) \, d\xi_i \, d\eta$$

measurement error — regression — population distribution of covariate

Lecture 11

17.2.25

(slides)

= outcome given overlap: multiply marginal likelihood to get likelihood for all data

Observed data likelihood:

$$P(x,y|\theta,\psi) = \prod_{i=1}^{N} P(x_i,y_i|\theta,\psi) \qquad \text{(indep.)}$$

In frequentist statistics: distinction between data and parameters: parameters are fixed and unknown, but not "random". Only "data" are random realisations of random variables.

What is the nature of the latent variables ξ_i, η_i?

- They have distribution $(\xi_i,\eta_i) \sim P(\xi_i,\eta_i|\theta,\psi) = P(\eta_i|\xi_i,\theta)\,P(\xi_i|\psi)$
- Often called "nuisance parameters" → needed to complete the model but not the parameters of interest (θ,ψ)
- Are the latent variables "data" or "parameters"?

("missing data" / "nuisance parameters")

BAYESIAN VIEWPOINT

* There is a symmetry between data D and parameters θ - both are r.v.s described by probability distributions
* Actually they are described by a joint probability

$$P(D,\theta)$$

(slides)

* Data are r.v.s whose realisations are observed, parameters are r.v.s not observed

* Goal is to infer the unobserved parameters from the observed data using the rules of probability

$$P(\theta|D) = \frac{P(D,\theta)}{P(D)}$$ CONDITIONAL PROBABILITY

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$ BAYES' THEOREM

* Probability interpreted as degree of belief/uncertainty in hypothesis — how sure are you in the values of that parameter?

Bayes' Theorem

- Joint probability of data & parameters: $P(D,\theta) = P(D|\theta)P(\theta) = P(\theta|D)P(D)$

- Probability of parameters given data:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(\theta|D)$: posterior prob.: degree of belief

$P(D|\theta)$: likelihood / sampling distribution

$P(\theta)$: prior prob.: degree of belief

$P(D)$: normalisation constant

example: bayesian inference of the dimensionality of space time (see lecture)

observe light & detect w interferometer
signal from black hole merger (?)
can use info from both to infer dimensionality of spacetime

dimensionality inference how long signal takes to reach us → can compare with how long light took to reach us

Simple Gaussian Example (slides)

Frequentist Confidence vs. Bayesian Credible intervals

(1) Frequentist Confidence Interval

$Y_1, \ldots, Y_4 \overset{iid}{\sim} N(\mu, 1)$

Sampling dist. of statistic (sample mean) $\bar{Y} \sim N(\mu, \sigma^2/4)$

$[\bar{Y} - \sigma_{\bar{y}}, \bar{Y} + \sigma_{\bar{y}}]$ is a 68% confidence interval

* Under repeated experiments, 68% of the (random) confidence intervals constructed this way will contain (cover) μ

This does NOT mean that the probability is 68% that μ lies within the interval (evaluated with y_{obs})

frequentist view: μ is a fixed number, it does not have a prob dist. interval will either cover that number or it doesn't

(2) Bayesian credible Interval (board)

assume flat prior: $P(\mu) \propto 1$

every value of μ equally likely

not normalisable, cannot integrate $\pm\infty$ so say $\propto 1$ (*)

"improper prior"

can derive posterior $p(\mu | Y = y_{obs})$:

$Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$

unknown ↑ ↑ (known)

Likelihood: $P(\vec{y} | \mu, \sigma^2) = \prod_{i=1}^{N} N(y_i | \mu, \sigma^2)$ (multiply)

$$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - \mu)^2/\sigma^2}$$

$$= (2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2}$$

* Likelihood $P(\vec{y}|\mu,\sigma^2)$ only depends on data $\vec{y}$ through the sufficient statistics $\bar{y}$ and s^2.

* Any two datasets $\vec{y}'$ and $\vec{y}''$ with the same sufficient statistics should yield the same inferences

* Likelihood principle: all info about parameters from data is located in likelihood function

$$= (2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y}+\bar{y}-\mu)^2}$$

$$= (2\pi\sigma^2)^{-N/2} e^{-1/2\sigma^2\sum_{i=1}^{N}\left[(y_i-\bar{y})^2+2(\bar{y}-\mu)(y_i-\bar{y})+(\bar{y}-\mu)^2\right]}$$ ($2(\bar{y}-\mu)(y_i-\bar{y})$ sums to 0)

$$= (2\pi\sigma^2)^{-N/2} e^{-1/2\sigma^2\left[\sum_{i=1}^{N}(y_i-\bar{y})^2+N(\bar{y}-\mu)^2\right]}$$

$$P(\vec{y}|\mu,\sigma^2) = (2\pi\sigma^2)^{-N/2} e^{-\frac{(N-1)}{2\sigma^2}s^2} e^{-N/2\sigma^2(\bar{y}-\mu)^2}$$

sufficient statistics

where (def):

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{→ "simple mean"}$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2.$$

→ "sample variance" (unbiased estimator for σ^2)

Case 1: $\sigma^2 = 1$ is known, $P(\mu) \propto 1$ improper prior

Bayes' theorem $P(\mu|\vec{y}) \propto P(\vec{y}|\mu)P(\mu)$

(ignore constant terms)

$$P(\mu|\vec{y}) \propto e^{-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2}$$ (unnormalised posterior)

notice looks like part of a gaussian, only not given variable is μ so this must define probability dist. over μ

could: (integrate out μ)

$$P(\mu|\vec{y}) = Ae^{-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2} \rightarrow \int d\mu \rightarrow \text{Find } A$$

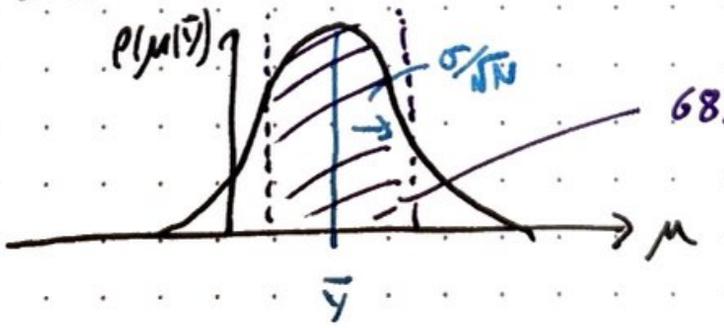
recognise gaussian

or:

$$P(\mu|\vec{y}) = \frac{1}{\sigma/\sqrt{N}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\bar{y}-\mu)^2}{(\sigma^2/N)}} \qquad \sigma^2/N$$

$$= N(\mu|\bar{y}, \sigma^2/N)$$ normalised posterior

gaussian on μ: tells us what degree of belief we have on various values of μ



integrate within purple lines gives

68% posterior probability

68% prob μ lies between $-x$ & $+x$

aka "credible interval"

~~See slide~~: frequentist vs bayes comparison for more detailed written overleaf:

Reading: ch 1-3 Gelman "Bayesian data analysis"
F & B 3.8
TVT 2.6.5

Lecture 12 (slides) 19.2.25

Frequentist vs. Bayes

* Frequentists make statements about the data (or statistics or estimators = functions of the data) conditional on the parameter
$$P(D|\theta) \text{ or } P(g(D)|\theta)$$

* Often goal is to get a "point estimate" or confidence intervals with good properties under repeated experiments $\to \infty$ ("long run")

* Arguments are based on datasets that could've happened but didn't e.g. null hypothesis testing

* Bayesians make statements about the probability of parameters conditional on the dataset $D = D_{obs}$ that you actually observed:
$$P(\theta | D = D_{obs})$$
This requires an interpretation of the probability as quantifying a "degree of belief" in a hypothesis

* Bayesian answer is the full posterior density
$$P(\theta | D = D_{obs})$$
quantifying the "state of knowledge" after seeing the data. Any numerical estimates are attempts to (imperfectly) summarise the posterior.

Bayes Advantages (slides)

* Ability to include prior information $P(\theta)$
  - Incorporate info. from external datasets. $P(\theta)$ is the posterior from some other data $P(\theta|D_{ext})$
    e.g. cosmology. diff datasets give diff ways to probe cosmological parameters
    ↳ encode external posterior from one analysis as prior to diff analysis. (natural way to incorporate information since one parameter often has multiple observables in cosmology)
  - Regularisation: penalises overfitting with complex model e.g. gaussian process prior
    (in frequentist land = have regularisation)
    e.g. linear regression. ridge regression is linear regression w/ a penalty term; in bayesian land we can incorporate such penalties. e.g. can say i don't think large slopes will occur and can incorporate this into prior. can also penalise complex models → Occam's razor.
    (e.g. flat; e.g. wide gaussian (slightly more info than flat))
  - "Noninformative" / weakly informative / default priors when you don't have much prior information
    e.g. might only be able to guess param is between 0 and 1 but still information! → cannot have prior without some information

(reminder: definition of likelihood in this course is diff. to how I usually think of it)

* Likelihood (you're plugged in observed dataset and are viewing that as fn. of the parameters) is not a probability density in the parameters. But multiply by a prior (even flat) and the (normalised) posterior is a probability density, conditional / marginal probabilities can be computed.
  (↳ e.g. in multi-parameter cases. useful to derive conditional densities (example later).

* Ability to deal with high dimensional parameter space e.g. (many) latent variables or nuisance parameters and marginalise them out (analytically/numerically)
  ↳ e.g. equivalent of MLE is called maximum a posteriori → can find peak of posterior and just use that as your point estimate

* Note: Bayesian inference not necessarily completely opposed to frequentist statistics. Estimators derived from Bayesian arguments can still be evaluated in a frequentist basis (e.g. James-Stein estimators)
  (can think of bayesian inference as just a way to generate estimators)
  (or find 95% credible intervals as intervals that contain some as e.g. posterior prob estimate)

(Last time: gaussian example of bayesian inference (1D, flat prior → like "hello world" of bayesian inference) is now slightly more involved example

recap:
Simple Gaussian Example w/ conjugate prior (board)

$Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$ (μ unknown, σ^2 known)

Likelihood: $P(\vec{y}\,|\mu,\sigma^2) = \prod_{i=1}^{N} N(y_i|\mu,\sigma^2) = \cdots$

$$= (2\pi\sigma^2)^{-N/2} e^{-\frac{(N-1)s^2}{2\sigma^2}} e^{-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2}$$

Defined **sufficient statistics**:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{"sample mean"}$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2 \quad \text{"sample variance"}$$

all info in data is contained in the likelihood & likelihood depends on sufficient statistics (summaries of data). e.g. if two diff y vectors w/ same $\bar{y}$ & s^2 would contain same info about parameters, even if y vector is different

Conjugate Prior: $P(\mu) = N(\mu|\mu_0, \tau_0^2)$

(equivalently can write $\mu \sim N(\mu_0, \tau_0^2)$) ($\mu_0$: prior mean, τ_0^2: prior variance)

$$P(\mu) = \frac{1}{\tau_0\sqrt{2\pi}} e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2}$$

Posterior: $P(\mu|\vec{y}) \propto P(\vec{y}|\mu)\,P(\mu)$

only fns of μ so discard other terms without μ

$$\propto e^{-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2} e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2}$$

recognize ☺

[warm up ex! – product of gaussian densities]

$$P(\mu|\vec{y}) = N(\mu|\mu_N, \tau_N^2)$$

(μ_N: posterior mean, τ_N^2: posterior variance)

- think: really a prob density on μ everything is determined except μ, μ
- -> nice! normalizing const s.t. when integrated out μ we get 1 const.
- -> useful actually to see it as product of gaussians

(warm up sheet 1)

reminder: when you have product of gaussian densities, the precision = $\frac{1}{\text{variance}}$ of the resulting posterior gaussian = sum of precisions of individual prior gaussians.

$$\left[\frac{1}{\tau_N^2} = \frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right]$$

POSTERIOR PRECISION

gaussians $N(\mu_1, \sigma_1^2)$ $N(\mu_2, \sigma_2^2)$

add them: $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

multiply them: $N\left(\frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \frac{1}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}}\right)$ ← normalizing constant

Posterior precision = sum of precisions from likelihood & prior

reminder:

$$\left[\mu_N = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{N}{\sigma^2}\bar{y}}{1/\tau_0^2 + \frac{N}{\sigma^2}}\right]$$

POSTERIOR MEAN

(divide by weights to get mean)

Posterior mean = precision weighted average of means from likelihood & prior

(These formulas explicitly tell us how in bayesian inference prior & likelihood (mean & variance) are combined to give us)

For fixed prior τ_0, as $N \to$ large $\mu_N = \bar{y}$ (tends to same as for flat prior, likelihood)

$p(\mu|y) = N\left(\mu \mid \frac{y\tau_0^2 + \mu_0\sigma^2/N}{\sigma^2/N + \tau_0^2}, \frac{\sigma^2/N\,\tau_0^2}{\sigma^2/N + \tau_0^2}\right)$ large

$\tau_N^2 \to \sigma^2/N$

(Data $N \to$ large, likelihood dominates over prior)

basically rather resulting likelihood fn. except w/ interpretation of probability

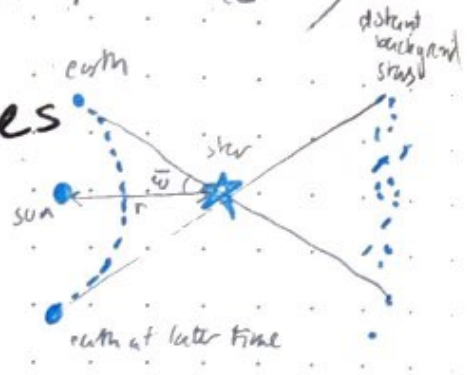
Side note: What is conjugate prior?
for a given likelihood if you combine it with a conjugate prior coming from 'nice' distribution then the posterior is guaranteed to come from the same class of 'nice' distributions (e.g. gaussian, wishart, laplace)

(not sure how relevant/necessary this is)

PAPER: (C Bailer-Jones Estimating distances from parallaxes)

Astrophysics Example (slides)

BAYESIAN INFERENCE FOR PARALLAX



- parallax is a way to measure distances

$$\frac{\omega_{true}}{\text{arcsec}} = \frac{\text{parsec}}{r}$$ (equivalently)

$$\omega_{true} = 1/r \quad \text{(in useful units)}$$

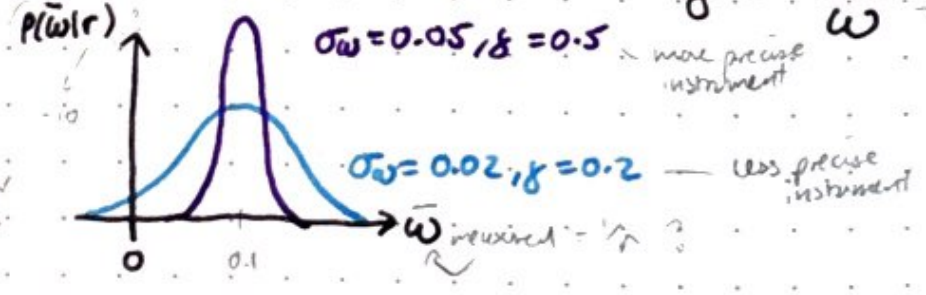
- Gaia satellite makes parallax measurement to measure distances & map stars in our galaxy

- measurement error in parallax angle measured parallax is different to true parallax

$$p(\bar{\omega}|r) = \frac{1}{\sigma_{\bar{\omega}}\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_{\bar{\omega}}^2}\left(\bar{\omega} - \frac{1}{r}\right)^2\right] \qquad \sigma_{\bar{\omega}} > 0$$

measured parallax has some distribution: assume gaussian distributed for simplicity

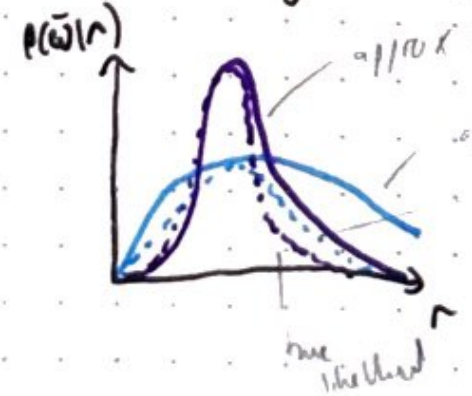
- fractional measurement error $f = \frac{\sigma_\omega}{\omega}$



r = 10pc:

if star at 10pc actual measured parallax would be one value drawn from this distribution

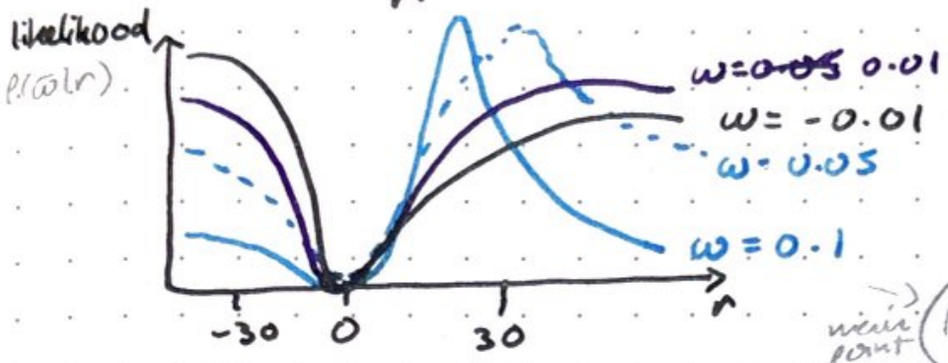
- possible to get negative parallax due to measurement error



likelihood

get skewed distributions

- Laplace approximation gives gaussian, incidentally this is mathematically equivalent to doing propagation of error (assume gaussian measurement error propagate to $\hat{r}$ ...)
  - this is an approximation (dotted vs solid line) ?



no peaks for -ve

Futility function

likelihood for various values of parallax

→ main point (plotting likelihood w/ no prior → gives junk results)

- likelihood is positive on negative values of distance (unphysical) – for all values of measured parallax
- negative measurements have no mode (MLE i.e. peaks → most likely distance peak for the but not for -ve

try to fix this:

- impose our knowledge that we know distans should be positive

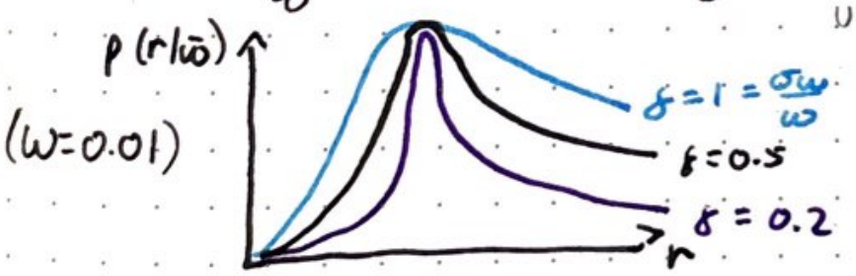
$$P(r|\omega) \propto P(\omega|r)\,P(r)$$

posterior

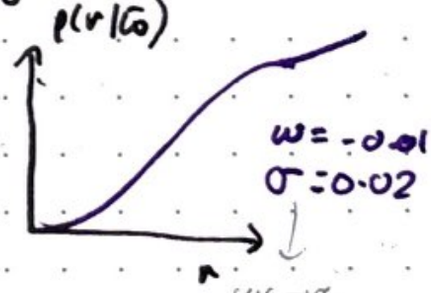
$$P(r) = \begin{cases} 1 & r>0 \\ 0 & \text{otherwise} \end{cases}$$

note: (improper distribution = not normalisable)

- this cuts off -ve values of r in our graph



unnormalised posterior



for -ve parallax, is not peaks at all → no mode

- improper posterior: not-normalisable no mean, variance etc.
- mode ($r = r_\omega$) exists for +ve ω but undefined for -ve ω ($r = \infty$)

still have some problems ↓

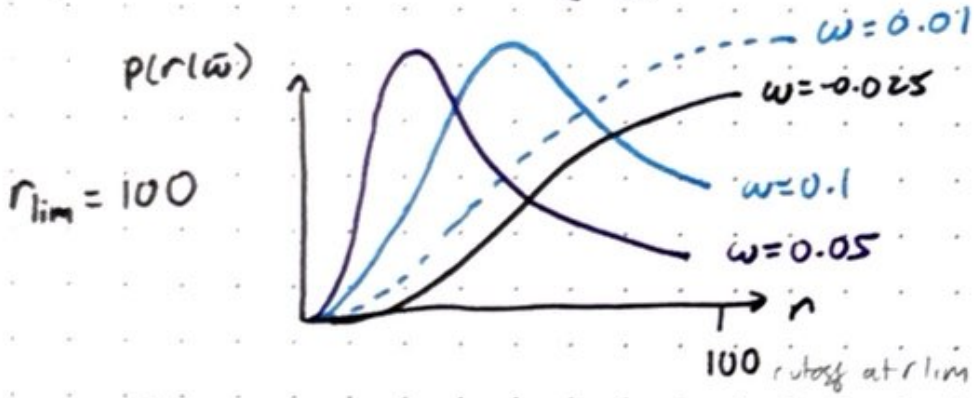
- impose a proper distance prior (impose limit to max distance of a star in our survey)

$$p(r) = \begin{cases} 1/r_{lim} & 0 < r \leq r_{lim} \\ 0 & \text{otherwise} \end{cases}$$

$$p(r|\omega) \propto p(\omega|r)p(r)$$

$$p_u^*(r|\bar{\omega}, \sigma_{\bar{\omega}}) = \begin{cases} 1/r_{lim}\, p(\bar{\omega}|r, \sigma_{\bar{\omega}}) & 0 < r \leq r_{lim} \\ 0 & \text{otherwise} \end{cases}$$

unnormalized posterior



- implies unrealistic stellar density $\sim 1/r^2$
- cutoff of mode at edge r_{lim} for small or negative measured parallax ($\omega < \sigma$) (but better than going to infinity)

$$r_{est} = \begin{cases} 1/\bar{\omega} & 0 < 1/\bar{\omega} \leq r_{lim} \\ r_{lim} & 1/\bar{\omega} > r_{lim} \quad \text{(extreme mode)} \\ r_{lim} & \bar{\omega} \text{ DNE} \leq 0 \end{cases}$$

- still have some problems → can go towards more & more astrophysically motivated priors

- problem w/ previous prior: volume density $p(r)$

$$Pr(\text{distance} \in [r, r+dr]) \propto p(r) 4\pi r^2 dr$$

unphysical →

$p(r) \propto 1/r^2$, $r < r_{lim}$; 0 otherwise $\Rightarrow$ $p(r) \propto$ const $r < r_{lim}$; 0 otherwise

- more physical: copernican principle consider $p(r) \propto$ const, $p(r) \propto$ const. $\times 4\pi r^2$

EE [illegible] same principle as previous priors

see slides for reading recs

Lecture 13

21.2.25

- astrophysics example
  recap: understanding how priors impact inference
- keeps going as before: introduce more & more astrophysically motivated priors (see slides)

Conclude: Priors in Bayesian Inference

* Priors can be used to encode background info./ external knowledge about parameters
  - weak mathematical constraints on physical parameters e.g. positivity of distances
  - astrophysical info. e.g. distributions of stars
* test sensitivity of your inference to the priors (& likelihood) under various assumptions of the model

$$P(\vec{w}|r) \propto P(r|\vec{w}) P(\vec{w})$$

$P(r|\vec{w})$ — measurement likelihood; $P(\vec{w})$ — prior; $P(\vec{w}|r)$ — posterior

note: also assumed some of measurement error e.g. gaussian not just prior

MULTIPARAMETEK BAYESIAN MODELS (slides)

Example: • gelman BDA 3.2 • sheet 2 [analytic posterior for gaussian, μ, σ^2 model with non-informative prior]

data generating process: $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$ $\quad i = 1, \dots, n$

likelihood: $p(y|\mu,\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right)$

sufficient statistics

(non informative) $p(\mu) \propto 1$

improper prior: $P(\log\sigma^2) \propto 1$ (ignorant of ordering mag. of (exponent of) σ^2) or $P(\sigma^2) \propto \sigma^{-2}$ JEFFREY'S PRIOR $(\sigma^2 > 0)$

even if prior improper, need proper posterior (corresponds to real observation)

joint posterior: $P(\mu,\sigma^2|y) \propto P(y|\mu,\sigma^2) \times P(\mu,\sigma^2)$

$$P(\mu,\sigma^2|\bar{y}) \propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right)$$

equivalent, showing change variable $P(\sigma^2) \propto \left|\frac{\partial \log\sigma^2}{\partial\sigma^2}\right| P(\log\sigma^2)$ $\times 1/\sigma^2$ segrey's prior

joint is useful to find conditional/marginals:

conditional posterior: $P(\mu|\sigma^2, y) = N(\mu|\bar{y}, \sigma^2/n)$

marginal posteriors: $P(\sigma^2|y) = \int P(\mu,\sigma^2|y)\,d\mu = \text{Inv-}\chi^2(\sigma^2|n-1, s^2)$ (ex sheet 2)

$$P(\mu|y) = \int P(\mu,\sigma^2|y)\,d\sigma^2$$

$$\propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}$$

$$= t_{n-1}(\mu|\bar{y}, s^2/n)$$ t distribution

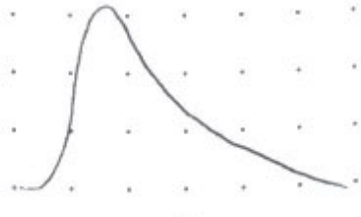
can derive these analytically is useful

Note: (1) **inverse chi-squared distribution**

$$\text{Inv-}\chi^2(\theta|\nu, s^2) \propto \theta^{-(\nu/2+1)} \exp\left(-\frac{\nu s^2}{2\theta}\right), \quad \theta > 0$$

deg. of freedom $\nu > 0$

scale parameter $s > 0$



(2)
same as inverse-gamma

$$\text{Inv-gamma}(\theta|\alpha,\beta) \propto \theta^{-(\alpha+1)} \exp(-\beta/\theta) \quad , \theta > 0$$

shape parameter $\alpha = \nu_0/2$
scale parameter $\beta = \frac{\nu_0}{2}\sigma_0^2$

useful: named distributions have random number generators in common python packages

(3)
student's t distribution

$$t_\nu(\theta|\mu,\sigma^2) \propto \left[1 + \frac{1}{\nu}\left(\frac{\theta-\mu^2}{\sigma^2}\right)^2\right]^{-(\nu+1)/2}$$

deg. of freedom $\nu > 0$
location parameter μ
scale parameter $\sigma > 0$

details on distributions in book (Gelman BDA) appendix
'or your own examples sheet' "or just wikipedia"

What if you can't compute marginals/expectations analytically? → **Bayesian computation**

* Bayesian answer: full posterior $p(\theta|D)$, numerical estimates are attempts to (imperfectly) summarise the posterior e.g. mean, mode

* Often these are posterior expectations $\mathbb{E}[g(\theta)|D)]$ (computationally difficult) $= \int g(\theta) p(\theta|D) d\theta$

* Bayesian computation: algorithms to "map out" / sample the posterior $p(\theta|D)$ and compute expectations $\mathbb{E}[g|\theta|D)]$

* e.g. MCMC, nested sampling, importance sampling

all models are wrong, some are useful (even if your computation is 'right'!)

object of computation is to calculate some integral (posterior expectation) → can do with monte carlo integration

MONTE CARLO INTEGRATION (slides)

Typically we want to summarize the posterior & compute expectations of the form

$$I = \mathbb{E}[f(\theta)|D] = \int f(\theta) p(\theta|D) d\theta$$

Using m samples from the posterior

$$\theta_i \sim p(\theta|D)$$

$$\hat{I} = \frac{1}{m}\sum_{i=1}^{m} f(\theta_i) \to I$$ (LLN for large N) /CLT

Monte Carlo Error (derive later on next page)

$$\mathrm{Var}[\hat{I}] = \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}[f(\theta)] = \frac{1}{m}\mathrm{Var}[f(\theta)] \approx \frac{1}{m}\widehat{\mathrm{Var}}[\{f(\theta_i)\}]$$

Monte carlo error $\sim 1/m^{1/2}$ indep of $\dim[\theta]$ → convenient

Fundamental Theorem of Monte Carlo → Bayesian computation using sampling

$$\underbrace{\mathbb{E}[f(\theta)|D]}_{\text{posterior expectation}} = \int f(\theta) p(\theta|D) d\theta \approx \underbrace{\frac{1}{m}\sum_{i=1}^{m} f(\theta_i)}_{\text{sample average}}$$

E.g.

Posterior mean μ	$f(\theta) = \theta$
Posterior variance	$f(\theta) = (\theta - \mu)^2$
Probability in an interval $[a,b]$	$f(\theta) = I_{[a,b]}(\theta)$

(can only do this if μ known?)

↑ (indicator function) = one when variable between a & b.

Monte Carlo Error (board)

suppose: f is some invertible fn. of θ

$\theta_i \overset{iid}{\sim} P(\theta)$ $\quad i = 1, \dots, m$ — monte carlo sample size (not same as data sample size)

→ $f(\theta_i)$ are iid

Def: $\hat{I} = \frac{1}{m} \sum_{i=1}^{m} f(\theta_i)$

$$\text{Var}(\hat{I}) = \text{Cov}(\hat{I}, \hat{I}) = \text{Cov}\left[\frac{1}{m}\sum_{i=1}^{m} f(\theta_i), \frac{1}{m}\sum_{j=1}^{m} f(\theta_j)\right]$$

$$= \frac{1}{m^2} \sum_{i=1}^{m}\sum_{j=1}^{m} \text{Cov}[f(\theta_i), f(\theta_j)]$$

θ_i, θ_j iid draws => only non-zero contributions when $j = i$

$$= \frac{1}{m^2} \sum_{i=1}^{m} \text{Var}(f(\theta_i))$$

$$= \frac{1}{m} \text{Var}(f(\theta))$$

($\frac{1}{m}$ Var $[f(\theta)]$)

standard deviation of monte carlo error

std MC error $\propto \frac{1}{\sqrt{m}}$

f

GLT: $m \to$ large $\quad \hat{I} \sim N\left(E[f(\theta)], \text{Var}(F(\theta))/m\right)$

slow convergence (increasing m reduces MC error)

but! → nice: indep. of dim(θ)

so MC error rate is indep of dimensions ☺ useful!

approximate variance

$$\text{Var}(\hat{I}) \approx \frac{1}{m} \widehat{\text{Var}}\left(\{F(\theta_i)\}\right)$$

↑ MC sample variance

$$\widehat{\text{Var}}\left[\{F(\theta_i)\}\right] = \frac{1}{m-1} \sum_{i=1}^{m} \left(f(\theta_i) - \hat{I}\right)^2$$

- MC error scales as $m^{-1/2}$
- Slow but indep of dim(θ)!
- Can approx

Lecture 14

notes on useful properties of gaussians on moodle!
↳ for ex sheet 2

24.2.25

Today: Bayesian Computation, Direct sampling, Importance sampling

Monte carlo direct sampling (slides)

example from before
↳ gaussian, μ, σ^2 modelled with non-informative prior
↳ $P(\mu) \propto 1$

how do we sample from joint? -> it factorises

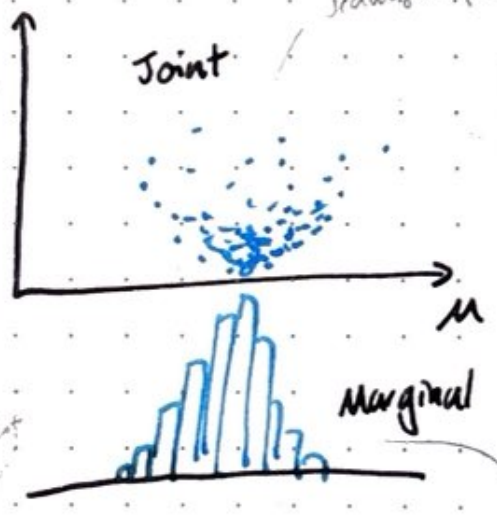
Factorise posterior: $P(\mu, \sigma | y) = P(\mu | \sigma^2, y) P(\sigma^2 | y)$

1. $\sigma_i^2 \sim P(\sigma^2 | y)$ — Inv-χ^2
2. $\mu_i | \sigma_i^2 \sim P(\mu | \sigma_i^2, y)$ — Normal

→ $(\mu_i, \sigma_i^2) \sim P(\mu, \sigma^2 | y)$

Joint draw from posterior!

draws from posterior



$g=0$
$s^2=1$
$n=10$
e.g. arbitrary choice of sufficient statistics

"removes summaries of data, all info about data contained in sufficient statistics"

Compute posterior summaries from monte carlo!

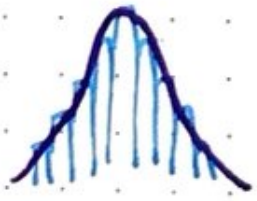
$E[\sigma|y] = 1.10$
$\text{std}[\sigma|y] = 0.25$

$E[\mu|y] = 0.00$
$\text{std}[\mu|y] = 0.14$

↳ "t sure how this rule"

(did not need to solve analytically to get marginals)

Kernel density estimate = estimate a smooth density from samples



"can think of as a smoothed histogram"

KERNEL DENSITY ESTIMATION (Laplace)

$\theta_i \overset{iid}{\sim} P(\theta|D) \quad i = 1, \ldots, m$

Approx $P(\theta|D)$ using $\hat{P}(\theta|D) \approx \frac{1}{m}\sum_{i=1}^{m} N(\theta|\theta_i; bw^2)$

$\int \hat{P}(\theta|D)\, d\theta = 1$ ✓ nice! normalized b/c gaussians sum to 1

select bandwidth

Lots of ways to choose bandwidth, nice one is

Silverman's Rule of Thumb:

$$bw = \left(\frac{4\hat{\sigma}^5}{3m}\right)^{1/5} \qquad \hat{\sigma}^2 = Var(\theta|D) \approx \widehat{Var}(\{\theta_i\})$$

sample variance $\frac{1}{m-1}\sum_{i=1}^{m}(\theta_i - \bar{\theta})^2$?

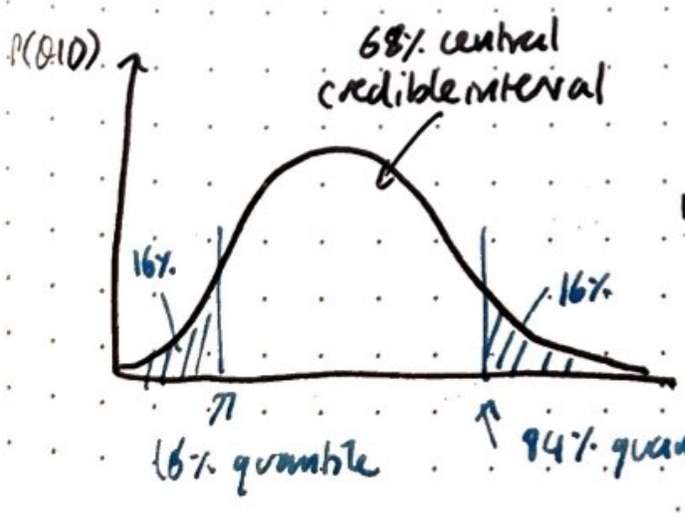
SUMMARISING POSTERIOR UNCERTAINTIES

* Can compute posterior mean (or median) ± posterior standard deviation
* If posterior $P(\theta|D)$ is Gaussian, then this contains 68% posterior probability
* But not necessarily for non-gaussian posterior!

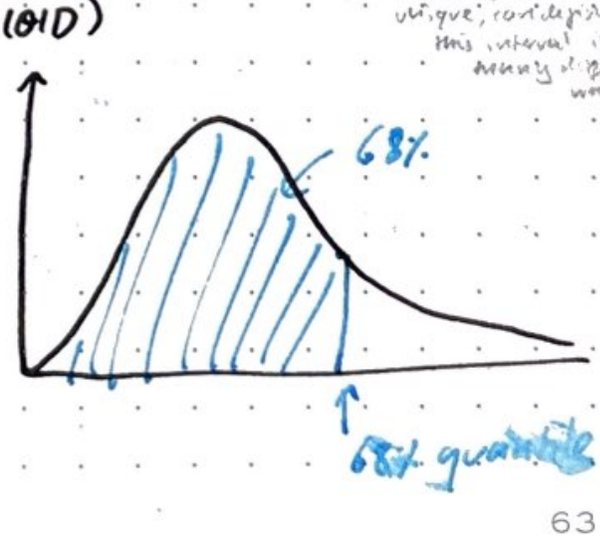
CREDIBLE INTERVALS

"In Bayesian land we talk about credible intervals"

prob θ lies in given interval is 68% → not unique, consider this interval in many diff ways



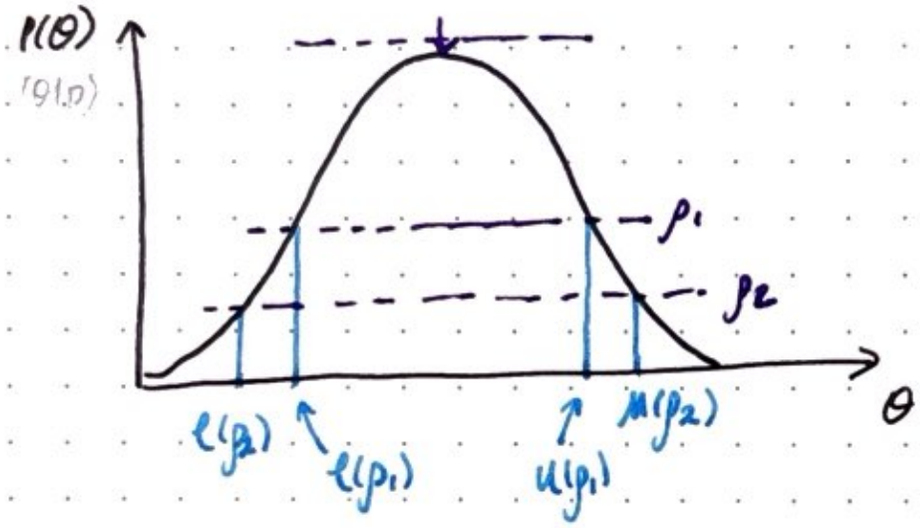
NOT UNIQUE!



(HPD) (board)

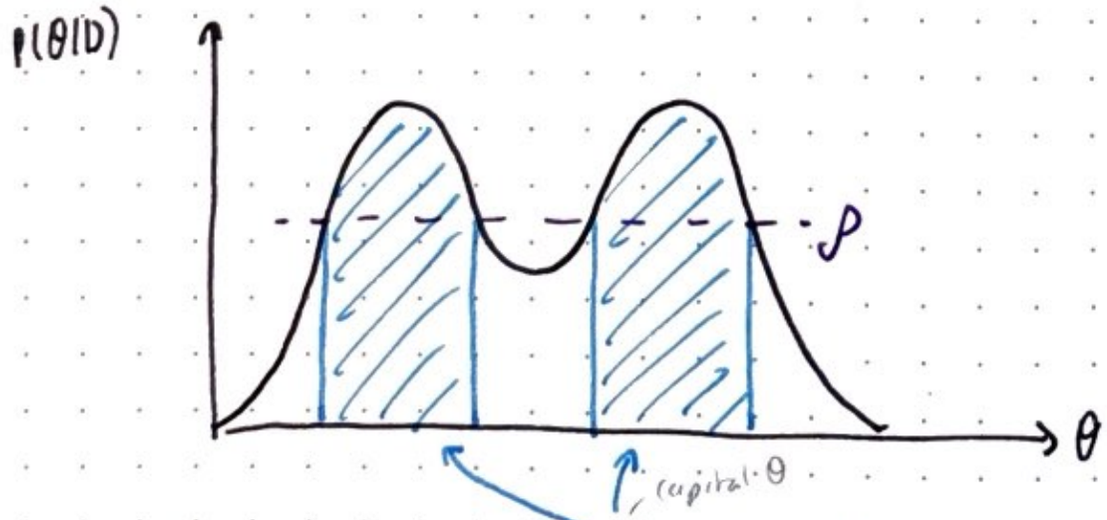
HIGHEST POSTERIOR DENSITY CREDIBLE INTERVAL

- Unique
- Narrowest interval that contains X% posterior probability



Find ρ s.t. $[\ell(\rho), u(\rho)]$ contains X% posterior probability.

HPD works for multi-modal posteriors!



X% HPD-C.I. = set Θ (possibly disconnected) with highest ρ s.t.

$\forall \theta \in \Theta$, $P(\theta|D) \geq \rho$ and $\int_{\Theta} P(\theta|D) = X$

What if you can't directly sample the posterior $\theta_i \sim P(\theta|D)$? (bound)

$$\mathbb{E}[g(\theta)|D] = \int g(\theta) P(\theta|D) d\theta \approx \frac{1}{m} \sum_{i=1}^{m} g(\theta_i)$$

e.g. likelihood might be output of some compute program

* Importance sampling:

  draw from an easier "tractable" distribution (importance function) $\theta_i \sim Q(\theta)$ and weight the samples by $w_i = P(\theta_i|D)/Q(\theta_i)$ to compute expectations

* Posterior simulation:

  MCMC, Nested sampling etc. generates draws from the posterior density iteratively in long-run

$$P(\theta|D) = \frac{P(D|\theta) P(\theta)}{P(D)}$$

IMPORTANCE SAMPLING (board)

Given a probability density $P(\theta)$ [e.g. the posterior]

Want to estimate

$$I = \mathbb{E}_p[g(\theta)] = \int g(\vec{\theta}) P(\vec{\theta}) d\vec{\theta}$$

However, $P(\vec{\theta})$ is intractable, i.e. difficult to directly sample, e.g. $\theta_i \sim P(\theta)$, but can evaluate $P(\theta_i)$ for a parameter value θ_i.

Select an instrumental distribution (importance function) from which it is easy to draw samples:

$$\theta_i \sim Q(\theta)$$ → tractible can easily draw/sample → evaluate

and can evaluate $Q(\theta_i)$, and $Q(\theta) > 0$ wherever $P(\theta) > 0$.

Rewrite $$I = \mathbb{E}_p[g(\theta)] = \int g(\theta) \frac{P(\theta)}{Q(\theta)} Q(\theta) d\theta$$

$$= \mathbb{E}_Q\left[g(\theta) \frac{P(\theta)}{Q(\theta)}\right]$$

Draw samples $\theta_i \overset{iid}{\sim} Q(\theta)$ $(i = 1, \ldots, m)$

Approximate I with importance sampling estimate:

$$\hat{I}^* = \frac{1}{m} \sum_{i=1}^{m} g(\theta_i) \frac{P(\theta_i)}{Q(\theta_i)} = \frac{1}{m} \sum_{i=1}^{m} g(\theta_i) w_i^*$$

where $w_i^* = w^*(\theta_i)$ = importance weights = $P(\theta_i)/Q(\theta_i)$

$$\mathbb{E}_Q[\hat{I}^*] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_Q[f(\theta_i)w^*(\theta_i)]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_Q\left[f(\theta_i)\frac{p(\theta_i)}{Q(\theta_i)}\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \int f(\theta)\frac{p(\theta)}{Q(\theta)}Q(\theta)\,d\theta$$

$$= \frac{1}{m}\sum_{i=1}^{m} \int f(\theta)p(\theta)\,d\theta$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_p[f(\theta)] = \mathbb{E}_p[f(\theta)]$$

$$= I \quad \text{(unbiased)}$$

reminder:
$\mathrm{Var}(aX+bY)$
$= a^2\mathrm{Var}X + b^2\mathrm{Var}Y$

$$\mathrm{Var}_Q[\hat{I}^*] = \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}\left(f(\theta)w^*(\theta)\right)$$

$$= \frac{1}{m}\mathrm{Var}\left(f(\theta)w^*(\theta)\right)$$

$$\approx \frac{1}{m}\hat{\mathrm{Var}}\left(\{f(\theta_i)w^*(\theta_i)\}\right)$$

$$\hat{\mathrm{Var}}(\{\}) = \frac{1}{m-1}\sum_{i=1}^{m}\left(f(\theta_i)w^*(\theta_i) - \hat{I}^*\right)^2$$

Lecture 15

useful properties of gaussians notes on moodle see sheet 2

26-2-25

Today: case study "Importance sampling the mass of the milky way galaxy"
Continuing: Bayesian computation, importance sampling MCMC (Gelman BDA ch10-12)

Recap: MC integration, want to summarize posterior, estimate w/ samples from the posterior. (replace integral w/ MC sum) which converges to desired expectation (summary statistic) for large n. Error on estimate $\propto 1/\sqrt{n}$ and is indep of dim of parameter space!

Importance sampling: recap (slides)

Objective: compute expectation w.r.t. distribution $p(\theta)$

$$I = \mathbb{E}[g(\theta)] = \int g(\theta) p(\theta) d\theta$$

Example: Posterior (suppress "1D" here)

$$p(\theta) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta) d\theta}$$

$L(\theta)$: likelihood, $\pi(\theta)$: prior

Can evaluate $p(\theta)$ directly but not sample from it. Choose importance function $Q(\theta)$ you can evaluate and sample!

$Q(\theta)$: named distribution; evaluate probability

Importance sampling estimate $\theta_1, \dots, \theta_n \overset{iid}{\sim} Q(\theta)$

$$\hat{I}^* = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i) w^*(\theta_i)$$

$$w_i^* = w^*(\theta_i) = p(\theta_i)/Q(\theta_i)$$

SELF-NORMALISED IMPORTANCE SAMPLING (slides)

Objective: compute expectation

$$I = \mathbb{E}[f(\theta)] = \int f(\theta) P(\theta) d\theta$$

but can only evaluate unnormalised $\tilde{P}(\theta)$.

Example: unnormalised posterior

$$\tilde{P}(\theta) = L(\theta)\pi(\theta).$$

(evidence: hard to evaluate)

Normalised posterior is $P(\theta) = \frac{\tilde{P}(\theta)}{Z_p} = \frac{\tilde{P}(\theta)}{\int \tilde{P}(\theta) d\theta}$,

but, cannot easily calculate

$Z_p = \int L(\theta)\pi(\theta) d\theta$ so cannot evaluate $P(\theta)$.

self-normalised importance sampling
aka SNIS

Self-Normalised Importance Weighting (hoped)

Often in Bayesian Analysis, you can only computer evaluate posterior $p(\theta)$ up to a constant.

$$P(\theta) = \frac{L(\theta)\pi(\theta)}{Z_p} = \frac{\hat{P}(\theta)}{Z_p}, \quad Z_p = \int \hat{P}(\theta)\, d\theta$$

($P(\theta)$: normalised; $\hat{P}(\theta)$: unnormalised)

GOAL: Estimate

$$I = \mathbb{E}_p[f(\theta)] = \int f(\theta) P(\theta)\, d\theta$$

$$= \int f(\theta) \frac{\hat{P}(\theta)}{\int \hat{P}(\theta) d\theta}\, d\theta$$

$$= \frac{\int f(\theta) \frac{\hat{P}(\theta)}{Q(\theta)} Q(\theta)\, d\theta}{\int \frac{\hat{P}(\theta)}{Q(\theta)} Q(\theta) d\theta}$$

"if you can't solve a hard problem, solve an easy problem"

instrumental distribution — preferably nice named distribution

Choose Q s.t. drawing samples $\theta_i \overset{iid}{\sim} Q(\theta)$ is easy $(i = 1, \dots, m)$.

don't need to solve for top & bottom integral?

Approximate I with (FTMC)

$$I \approx \hat{I} = \frac{\sum_{i=1}^{m} f(\theta_i) \hat{w}(\theta_i)}{\sum_{i=1}^{m} \hat{w}(\theta_i)} = \sum_{i=1}^{m} f(\theta_i) w(\theta_i)$$

(Estimate normalisation using same samples from Q)

where the self-normalised weights are,

$$w(\theta_i) = \frac{\hat{w}(\theta_i)}{\sum_{i=1}^{m} \hat{w}(\theta_i)} = \frac{\hat{P}(\theta_i)/Q(\theta_i)}{\sum_{i=1}^{m} \hat{P}(\theta_i)/Q(\theta_i)}$$

nice: works when $P(\cdot)$ unnormalised.
we have done the bottom integral using importance sampling itself.

(Also works if Q is unnormalised, since const. factor in Q cancel above)

COST: $\hat{I}$ is slightly biased (whereas importance sampling we looked at before unbiased)

$$\mathbb{E}[\hat{I}] = I + O(1/m)$$

from estimating denominator

b/c of denominator (work out yourself)

biased but consistent? → as $M \to \infty$ $P(\hat{I}=I) \to 1$

bias arises from self-normalisation process

reminder: $\mathbb{E}[AB] = E[A]E[B] + \text{Cov}(A,B)$
since $\text{Cov}(A,B) = E[AB] - E[A]E[B]$

cheat GPT attempt:

$$\mathbb{E}_Q[\hat{I}] = \sum_{i=1}^{m} \mathbb{E}_Q[g(\theta)w(\theta)] = \sum_{i=1}^{m} \left(\mathbb{E}_Q[g(\theta)]\mathbb{E}_Q[w(\theta)] + \text{Cov}(g(\theta), w(\theta)) \right)$$

unclear steps:

$\mathbb{E}_Q[w(\theta)] = 1/m + O(1/m^2)$ because $\sum w_i = 1$?

(and $\text{Cov}(g(\theta), w(\theta))$ ignorable? $O(1/m^2)$?)

$$\rightarrow \mathbb{E}_Q[\hat{I}] = \sum_{i=1}^{m} \left(\mathbb{E}_Q[g(\theta)] \left(1/m + O(1/m^2) \right) \right)$$

$$= \mathbb{E}_Q[g(\theta)] + O(1/m)$$

some kind of perturbation analysis / taylor expansion?

$$\text{Var}[\hat{I}] = \sum_{i=1}^{m} \text{Var}[g(\theta_i)w(\theta_i)] = m\,\text{Var}[g(\theta)w(\theta)]$$

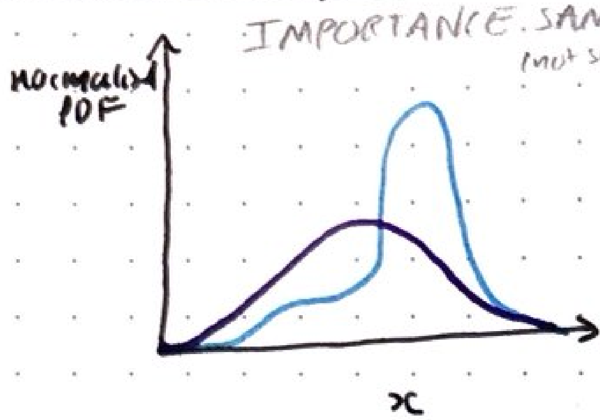
$$\approx m\,\widehat{\text{Var}}(\{g(\theta_i)w(\theta_i)\})$$?

? idk

(slides)

Contrived example: Gaussian Mixture, normalised PDF

IMPORTANCE SAMPLING EXAMPLE
(not self normalised)



mixture of 2 gaussians

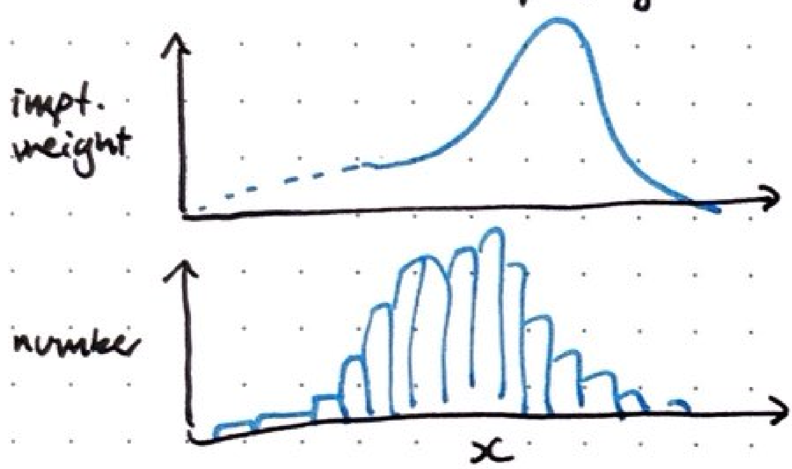
— target $P(x)$

— impt. fn. $Q(x)$ — we choose gaussian

target mean = 0.20

(have analytic pdf but suppose we didn't know this & try to estimate)

Generate 1000 samples from impt. fn.



MC estimate of mean

$\sum_m x_i w_i^* = 0.215 \pm 0.028$

(within 1σ of target)

m=1000
draws from impt. fn.

could also do this for posterior variance etc.

m=10,000 draws: MC estimate of mean is

$\sum_m x_i w_i^* = 0.199 \pm 0.009$ — error goes down as $1/\sqrt{m}$

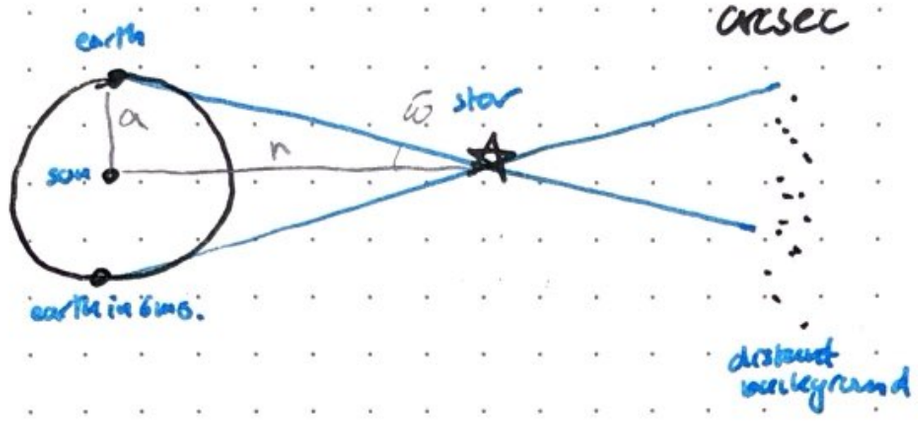
Parallax Example (slides)

Likelihood $\omega \sim N(1/r, \sigma_\omega^2)$ (gaussian measurement error in ω?)

$$P(\bar{\omega}|r) = \frac{1}{\sigma_{\bar{\omega}}\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_{\bar{\omega}}^2}\left(\bar{\omega} - \frac{1}{r}\right)^2\right], \quad \sigma_{\bar{\omega}} > 0$$

parallax recap

True relation (no errors): $\frac{\omega}{\text{arcsec}} = \frac{\text{parsec}}{r}$ ← parallax angle, ← distance



since $\bar{\omega} \ll 1$
- $\bar{\omega} = a/r$ to v. good approx
- when $\bar{\omega}$ is 1 arcsecond, r is defined as 1 parsec

Introduce physically motivated prior (as seen before some kind of)

$$P(r) = \begin{cases} \frac{1}{2L^3} r^2 e^{-r/L} & \text{if } r > 0 \\ 0 & \text{otherwise} \end{cases}$$

Exponential decrease in density of stars with galactic length scale L

$$P(r|\omega) \propto P(\omega|r)\,P(r)$$

Unnormalised posterior

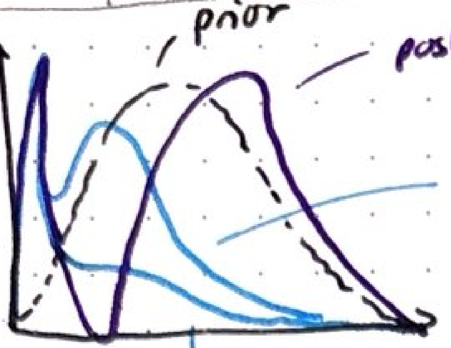
$$P^*_{r^2e^{-r}}(r|\bar{\omega}, \sigma_{\bar{\omega}}) = \begin{cases} \frac{r^2 e^{-r/L}}{\sigma_{\bar{\omega}}} \exp\left[-\frac{1}{2\sigma_{\bar{\omega}}^2}\left(\bar{\omega} - \frac{1}{r}\right)^2\right] & r > 0 \\ 0 & \text{otherwise} \end{cases}$$

can't compute normalisation analytically

With reality this is a 1D example so we could actually evaluate this on a grid and do integral numerically but let's pretend we can't do that

want to compute posterior mean: consider posteriors of distance

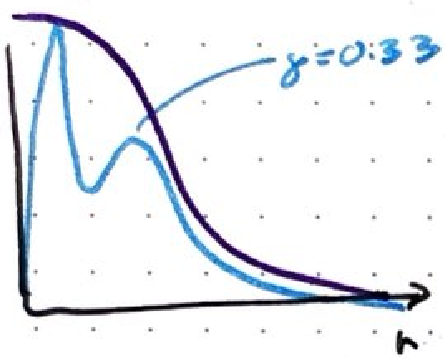
blue: posterior for $\bar{\omega} = 0.01$ with some value of error



short peak at low r ratio γ. bimodal for higher noise/signal. (arbitrary γ?)

[e.g. trying calc mean for $\omega = 0.01$ with error $\gamma = 0.33$) need to use self-normalized importance sampling => need to choose impt. fn.]

bigger the error the more it looks like prior



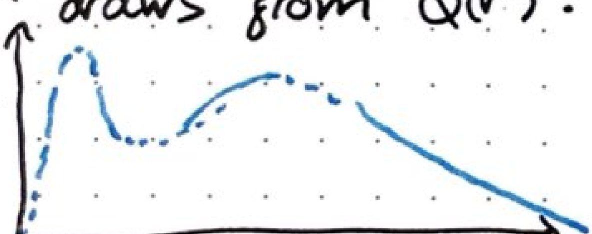
$\gamma = 0.33$
$\omega = 0.01$

— target posterior $P(r|\omega)$
— impt. fn. $Q(r)$

↳ choose half student t dist with $v=3$ (low deg of freedom)
↳ choose because tail of student t is thicker than gaussian (gaussian drops off as e^{-x^2} but student t e^{-x})
want $Q(r)$ to have fatter tail than the target, also choose b/c we want a named distribution

choose $Q(r)$ that covers $P(r)$ (target)
↳ should be +ve everywhere the target is +ve
(in this case choose half student's t) w/ low deg of freedom $v=3$

$M=10^4$ draws from $Q(r)$:

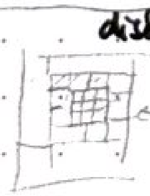


MC estimate of mean
$\sum x_i w_i = 2391.263 \pm 3.918$

$M=10000$ draws from impt. fn.

think of weights like adjusting resolution of grid! (density)

weights = shading



← optimal grid: low variance in weights

as evidenced by large variation in weights

↑ uniform: high variance in weights of each square

idea of ESS, can get same info from uniform grid as non uniform grid w fewer samples but with weights. so we get more effective info from each grid square → i.e. the uniform grid has low ESS

from article "a conceptual introduction to MCMC":

imagine approximating integral over prob dist as a sum over a grid of points. grid squares in low prob areas do not contribute a lot to the sum. sum is in fact dominated by small no of points with highest "weights" / probabilities. so we want our grid to be finer here but coarser in regions that don't contribute much → equiv. to concept of having similar weights everywhere! so small variance in weights

CHOOSING A GOOD IMPORTANCE FN. (slides)

(most important part of importance sampling)

within class of named distributions:

* Can be shown (sheet 2) that theoretically optimal (minimum variance) importance fn. is

$$Q^*(\theta) = \frac{|f(\theta)| P(\theta)}{\int |f(\theta)| P(\theta) d\theta}$$

high variance in importance weight → comparing estimates (→ small effective sample size)

* However, if we can't directly sample from $P(\theta)$, then we probably can't sample from $|f(\theta)| P(\theta)$

* Want to keep the importance weights roughly const., otherwise large variations in $P(\theta)/Q(\theta)$ will lead to high variance of estimate, smaller ESS

* Effective sample size

i.e. importance fn similar shape to dist. we are trying to sample

$$ESS = \frac{m}{1 + \hat{Var}[\{\omega^*(\theta_i)\}]}$$

e.g. as in parallax example

* In practice, find thick-tailed distribution $Q(\theta)$ that is positive everywhere and similar in shape to $|f(\theta)| P(\theta)$

* Don't want $Q(\theta)$ small when $|f(\theta)| P(\theta)$ large!

problem: small sample size, high variance.

→ ideal importance fn. is proportional to $|f(\theta)| P(\theta)$

→ closely resembles target pdf (want to have this) (i.e. importance weights ≈ const ⇒ small variance in weights ⇒ large ESS)

→ in practice want positive everywhere, covers the tail (thick tail), could (thick) maybe use/find/or named distributions

Intro (slides)

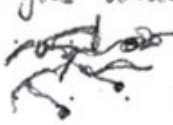
satellite galaxies; large magellanic cloud; milky way (central galaxy); SMC; Andromeda

Astrostatistics case study

"Bayesian estimates of the milky way and andromeda masses using high-precision astrometry & cosmological simulations"

- want to measure mass of milky way (constrain mass of our galaxy)
- satellite galaxies of our milky way (eg LMH.) -> use their orbits/motions to estimate mass of our own galaxy
- measure distance of satellite galaxy to central galaxy, angular momentum etc.
- use simulation to plot kinematic properties against mass of central galaxy (correlations between the two)
  - ↳ p(angular mom.| mass)
- use bayesian inference to infer actual mass of our galaxy from measured kinematic properties & results of simulations

illustris simulation: simulate evolution of galaxies
small density fluctuations overdensity regions attract mass, galaxies form (attract gas which forms stars -> galaxies.)



(web)

- can measure properties of satellite galaxy properties vs mass of central galaxy -> get catalogue of properties
- simulation generates samples (prior p(mass) p(angular mom.)) -> intractible in sense that we can't write it down but we can draw from it)

velocities (v) positions (r) momenta (j) of satellites correlated w/ central galaxy mass via galaxy formation simulations (physical prior) (intractable but can sample from prior)

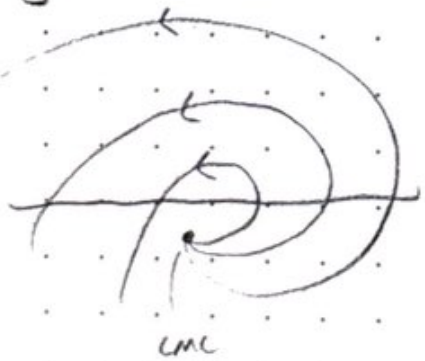
x = latent (true) values v, r, j M_{vir} = mass of galaxy, θ = params x, M_{vir}

(joint prior $p(v, r, j, M)$ from simulations)

measure LMC data v, r, j (from HST) we have analytic measurement likelihood $p(x|d)$, now use importance sampling to estimate posterior!

$$p(x|d) \propto L(x|d) = N(v_{obs}|v, \sigma_v^2) \times N(r_{obs}|r, \sigma_r^2) \times N(v_{obs}|v_{hst}, \sigma_{hst}^2)$$

-> Bayesian inference
-> Importance sampling
-> kernel density estimation



(board)

IMPORTANCE SAMPLING THE MILKY WAY GALAXY (Adel+17)

$\vec{d}$ = (noisy) measurements of kinematic properties of our Milky way satellite's (e.g. large Magellanic Cloud) properties (e.g. distance from central galaxy, velocity, angular momentum...) $\vec{x} = v_i, r_i, j$.

$\vec{x}$ = latent (true) values of satellite properties

$m = \log_{10}$ [mass of central galaxy]

GOAL: Estimate $\mathbb{E}[g(m)|\vec{d}] = \int g(m) P(m|\vec{d}) dm$

(e.g. let $g(m) = m$ for posterior mean or $g(m) = (m - \mathbb{E}[m|d])^2$ for posterior variance

Posterior mean: $\mathbb{E}[m|\vec{d}] = \int m \underbrace{P(m|\vec{d})}_{\text{marginalised posterior}} dm$

$$= \int m \left[\int \underbrace{P(\vec{x}, m|\vec{d})}_{\text{full posterior}} d\vec{x} \right] dm$$

$$P(\vec{x}, m|\vec{d}) = \frac{P(\vec{d}|\vec{x}) P(\vec{x}, m)}{\iint P(\vec{d}|\vec{x}) P(\vec{x}, m) d\vec{x} dm}$$

(Gaussian) measurement likelihood

Physical prior (encoded in cosmo. sims)

why not $P(x|m)$? reasonable assumption if I know true values of e.g. V then knowing m does not provide any additional info. on V

$$\mathbb{E}[m|\vec{d}] = \frac{\iint m P(\vec{d}|\vec{x}) P(\vec{x}, m) d\vec{x} dm}{\iint P(\vec{d}|\vec{x}) P(\vec{x}, m) d\vec{x} dm}$$

(bottom is a number so don't need to integrate again)

Cosmo. sims generate galactic systems that can be thought of as giving samples $(\vec{x}_j, m_j) \overset{iid}{\sim} P(\vec{x}, m)$ $j = 1, \dots, m$ from the physical prior (intractible).

i.e. expected mass of our galaxy

Aim: → want $\mathbb{E}[m|d]$ so need $P(m|d)$ which we get from marginalising $P(x, m|d)$.

→ we know $P(x, m)$ from sims so find $P(x, m|d) = \frac{P(d|x, m) P(x, m)}{P(d)} = \frac{P(d|x) P(x, m)}{\text{const.}}$

→ $P(d|x)$ is just meas. error

→ then integrate to get $P(m|d)$ and integrate again to get $\mathbb{E}[m|d]$.

→ use the prior itself as the importance function!

(where we can get away with intractible Q (and prior?) since they cancel?)

Apply Monte Carlo approx.: (tractable think?)

$$\mathbb{E}[m|d] \simeq \frac{\frac{1}{n}\sum_{j=1}^{n} m_j P(\vec{d}|\vec{x}_j)}{\frac{1}{n}\sum_{j=1}^{n} P(\vec{d}|\vec{x}_j)} = \sum_{j=1}^{n} m_j w_j$$

i.e. from sims we have samples from prior but not prior dist. itself.
↓
now we only need samples from prior to eval $\mathbb{E}(m|d)$ which we have!)

(Repeat for posterior variance)

Self-norm importance weight

$$w_j = \frac{P(\vec{d}|\vec{x}_j)}{\sum_{k=1}^{n} P(\vec{d}|\vec{x}_k)}$$

$$\sum_{j=1}^{n} w_j = 1$$

REMARKS

removes w_j taking unnormalised posterior to Q fn.; so what happened to the Q. fn. here? → essentially cancels out because we used prior as Q fn.

$$w_j \propto \frac{P(d|x)\,P(x,m)}{Q(x,m)}$$

cancel because encode prior as importance function itself

↓

means we get away with not being able to evaluate density $Q(x,m)$ itself (because we can only get samples from computer distribution) since it cancels with the prior!

$P(d|x,m) = P(d|x)$

$\vec{d}$ conditionally indep. of m given $\vec{x}$.

conditional independence assumption: everything about true dist'n measures contained in $\vec{x}$

- if knew the $\vec{x}$ value ran monte carlo all the measurements and that is no additional info contained in the mass

measuring distance or velocity satellite galaxy does not probe mass directly
↳ the only info on mass comes from varying in the prior fn.

Lecture 16

28.2.25

Today: continuing Bayesian computation, importance sampling, MCMC

recommend: paper Speagle J "A Conceptual Introduction to Markov chain Monte Carlo Methods"

why do we not just do grid sampling / importance sampling all the time?

WEIGHTED KERNEL DENSITY ESTIMATE (slides)

$$\text{Wkde}(\theta) = \sum_{s=1}^{m} w_s \times N(\theta | \theta_s, bw^2)$$

$$w_s = \frac{\tilde{p}(\theta_s)/q(\theta_s)}{\sum \tilde{p}(\theta_s)/q(\theta_s)}$$

w_s = normalised importance weights

bandwidth: Silverman's rule of thumb $bw = \left(\frac{4\sigma^5}{3n}\right)^{1/5}$

Estimate σ^2 from posterior Var[θ] estimate.

Use importance sampling effective sample size (ESS) for n.

$$\frac{m}{1 + \widehat{\text{Var}}[w_s]}$$

Ideal case:

if equal weights: $w_i = 1/m$ reduces to

$$\text{Kde}(\theta) = \sum_{s=1}^{m} \frac{1}{m} \times N(\theta | \theta_s, bw^2)$$

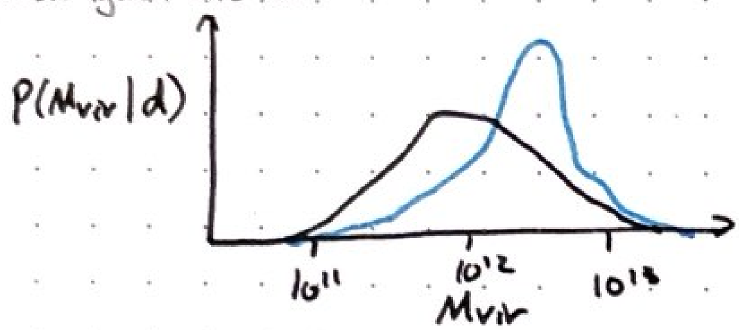
continue from example before:

basically "answer to sheet 2 q3"

Posterior of Milky Way w/ weighted KDE

(slides)

posterior given LHC data



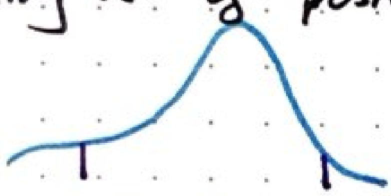
MW $M_{vir} = 1.7^{+1.33}_{-0.52} \times 10^{12}$

maximum weight??

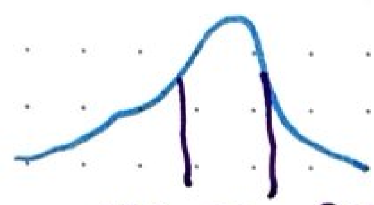
mean weight?

Posterior HPD

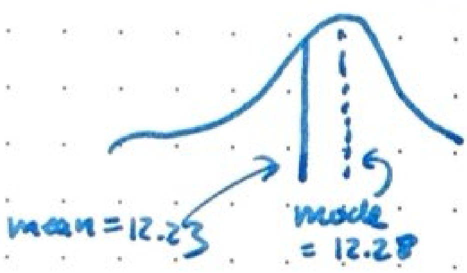
X% HPD (X% credible interval(s) with highest density containing X% of posterior)



90% HPD = [11.83, 12.65]



68% HPD = [12.13, 12.42]



can make combined posterior from data from multiple satellite galaxies & do same calculation w/ weighted KDE

combine inferences from different sources to get more accurate estimate (this is an ideal case for bayesian inference)

Frameshift: misconception that primary purpose of MCMC is to estimate the posterior, primary purpose is actually to estimate expectations!

MARKOV CHAIN MONTE CARLO MCMC

(slides)

class of algorithms for sampling from some probability dist.
→ useful for dists that are too complex or high-dimensional to study with analytic techniques alone (curse of dimensionality!)

Goal is to evaluate the posterior $P(\theta|D)$:

* Simple likelihoods / conjugate priors admit analytic solutions to the posterior
* Simple models may allow direct draws $\theta_i \sim P(\theta|D)$ i.e. "direct simulation"
* Small numbers of parameters p (small dimensionality)
  → Evaluate posterior on a p-dimensional grid.
  Inefficient for $p > 3$ (end up wasting compute time evaluating $P(\theta|D)$ where it is close to zero.
* Realistic models with many parameters p (high dimensional parameter space) (and we have no analytic solution)
  ↳ Markov Chain Monte Carlo is the gold standard!

whole

Point of MCMC is to generate samples from the posterior when we can't directly sample from it

$$\mathbb{E}[g(\theta)|D] = \int g(\theta) P(\theta|D) d\theta \approx \frac{1}{m} \sum_{i=1}^{m} g(\theta_i)$$

* Posterior simulation — MCMC
* How? - Generate a correlated sequence (chain) of random variates (monte carlo) that (in a limit) are draws from the posterior. The next value in the sequence only depends on the current values (Markov)

(explain how later)

* Algorithm cleverly constructed to ensure distribution of chain values → posterior dist. = stationary dist. in the long run.

Simplest MCMC: METROPOLIS ALGORITHM (slides)

(1) Choose random starting point μ_0

(2) At each step $i = 1, \ldots, N_{MC}$ propose a new parameter value μ_{prop} via $J(\mu)$. The proposal distribution $J(\mu)$ is usually $N(\mu | \mu_i, \tau^2)$

The proposal scale τ is chosen (cleverly!) for efficiency!

(3) Evaluate ratio of posteriors at proposed vs. current values

Metropolis Ratio $r = \frac{P(\mu_{prop} | y)}{P(\mu_i | y)}$

(4) If μ_{prop} is a better solution (higher posterior), $r > 1$, accept the new value

$$\mu_{i+1} = \mu_{prop}$$

Else accept with probability r (i.e. accept with probability $\min(r, 1)$). Stay at same value

$$\mu_{i+1} = \mu_i$$

(and include in chain - keep repeated values)

(5) Repeat 2-4 until reach some measure of convergence and gather enough samples to compute your inference.

Markov = next value depends on current value

Monte Carlo = randomness

return to example w/ analytic solution

Example: simple gaussian unknown μ (slides)

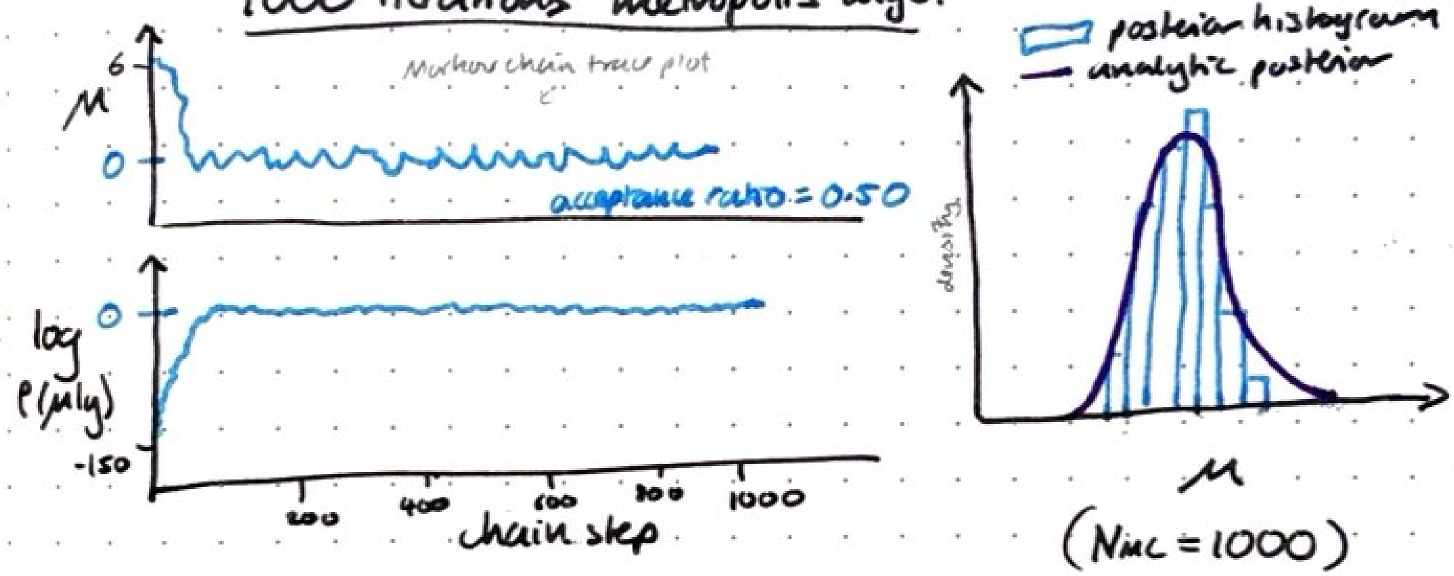
Likelihood: $y_i \sim N(\mu, \sigma^2 = 1) \quad i = 1, \dots, N$

Prior: $P(\mu) \propto 1$

$\Rightarrow$ Posterior: $P(\mu|y) = N(\mu|\bar{y}, \sigma^2/N)$ (analytic solution) (L17)

Do N_{mc} MCMC steps (proposal choose scale 0.6) track no. of acceptances
code on moodle to try yourself. choose $\bar{y}$ - (remember summary statistic contains all info about data so the posterior)

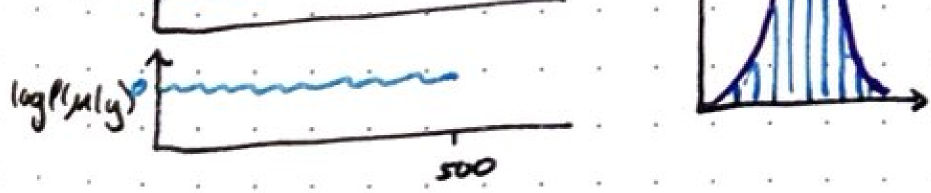
1000 iterations metropolis algo.:



During burn-in e.g. just 50 iterations acceptance ratio = 0.42

acceptance ratio goes up as if we make small steps r ratio goes up so accept more? idk what he's saying

Can also
Cut off 1st half of chain to cut burn in (aggressive solution, more empirical cut-offs exist)



(in reality if we had 1D dist. probably wouldn't need to do MCMC, could just give CDF, invert & apply to uniform rv to get draw from dist. but show here as example

(hint q5 sheet 2 on moodle)

Lecture 17

3.3.25

[Today: continuing Bayesian computation: MCMC]

Recap MCMC:

ideally want iid but in long run rely that we get ergodic indep so (?) that this works

not iid (correlated) but looks like iid for large N_{MC}

Generate a correlated sequence (chain) of random variates (Monte Carlo) that (in a limit) are draws from the posterior. The next value in the sequence only depends on current values (Markov).

D-DIM METROPOLIS ALGORITHM: (slides)

Posterior $P(\theta | D)$ where $\dim(\theta) = d$

Symmetric proposal/jump dist. $J(\theta^* | \theta) = J(\theta | \theta^*)$

1. Choose random starting point θ_0
2. At step $i = 1, \dots, N$ propose new parameter value

   $$\theta^* \sim N(\theta_{i-1}, \Sigma_p)$$

   (note to self: what is metropolis? symmetric jump (not necessarily normal) but knowing slides usually normal so assume metropolis is normal for exam?)

   The proposal distribution is

   $$J(\theta^* | \theta_{i-1}) = N(\theta^* | \theta_{i-1}, \Sigma_p)$$

3. Evaluate ratio of posteriors at proposed vs current values

   $$r = \frac{p(\theta^* | y)}{p(\theta_{i-1} | y)}$$

④ Accept θ^* with probability

$$\min(r, 1) : \theta_i = \theta^*$$

If not accept, stay at same value $\theta_i = \theta_{i-1}$ for the next step and include in chain.

⑤ Repeat 2-4 until reach some measure of convergence (G-R) and gather enough independent samples to compute your inference (reduce Monte Carlo error)

widely used way to sample from MVG:
[HOW DO WE DRAW FROM MVG?]

(i.e. step 2 d-dim metropolis)

Multivariate Gaussian Draws

$\vec{\theta} \in \mathbb{R}^d$

Metropolis proposal: $\vec{\theta}^* \sim N(\vec{\theta}_{i-1}, \underset{\sim}{\Sigma}_p)$

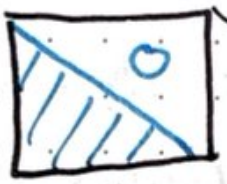
$\underset{\sim}{\Sigma}_p$ = real, symmetric, positive definite dxd matrix

$\det(\underset{\sim}{\Sigma}_p) > 0$ and $\underset{\sim}{\Sigma}_p^{-1}$ exists.

1) Cholesky decomposition, find $\underset{\sim}{L}$ s.t. $\underset{\sim}{\Sigma}_p = \underset{\sim}{L}\underset{\sim}{L}^T$ (i.e. in other words)

$\underset{\sim}{L}$ is lower triangular Cholesky factor

"matrix square root"



(cholesky decomp expensive so don't want to do cholesky decomp at every point in your loop. so instead do this)

2) Draw $\vec{\theta}^* \sim N(\vec{\theta}_{i-1}, \underset{\sim}{\Sigma}_p)$

by $\vec{\theta}^* = \vec{\theta}_{i-1} + \underset{\sim}{L}\vec{z}$ (do this computationally w/ code)

where $\vec{z}$ is a d-vector of iid ~~uniform~~ unit normal r.v.s

$$z_i \overset{iid}{\sim} N(0,1) \qquad i = 1, \ldots, d$$

check that sampling $\vec{\theta}^*$ like this is indeed the same as sampling from MVG:

Check: $\mathbb{E}[\vec{\theta}^*] = \vec{\theta}_{i-1}$

wiki: "$\vec{\theta}_{i-1} + \underset{\sim}{L}\vec{z}$ has the desired dist. due to the affine transformation property"

: any affine transformation $g(x) = Ax + b$ of a gaussian is a Gaussian

$$\mathrm{Cov}[\vec{\theta}^*, \vec{\theta}^{*T}] = \mathrm{Cov}[\underset{\sim}{L}\vec{z}, (\underset{\sim}{L}\vec{z})^T] \quad \text{(bilinear)}$$

$$= \underset{\sim}{L}\,\mathrm{Cov}[\vec{z}, \vec{z}^T\underset{\sim}{L}^T] = \underset{\sim}{L}\,\mathrm{Cov}[\vec{z}, \vec{z}^T]\,\underset{\sim}{L}^T$$

$$= \underset{\sim}{L}\,\underset{\sim}{I}\,\underset{\sim}{L}^T = \underset{\sim}{L}\underset{\sim}{L}^T = \underset{\sim}{\Sigma}_p$$

note to self: summary
1. resulting dist. gaussian due to affine transformation property
2. can show resulting gaussian has desired dependencies
3. and desired covariance matrix //

(slides)

Example:

Recall: analytic posterior density for Gaussian (μ, σ^2) model with non-informative priors $P(\mu) \propto 1$, $P(\sigma^2) \propto \sigma^{-2}$, $\sigma^2 > 0$. (multi-param. inference)

→ the mean interval?? identical to

* Joint posterior: $(\sigma^2 > 0)$

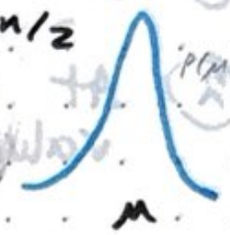
$$P(\mu, \sigma^2 | y) \propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right)$$

"marginal posterior"

* Marginal of μ:

$$P(\mu|y) = \int P(\mu, \sigma^2|y)\, d\sigma^2 \propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}$$

$$= t_{n-1}(\mu | \bar{y}, s^2/n) \quad \text{(t-distribution)}$$

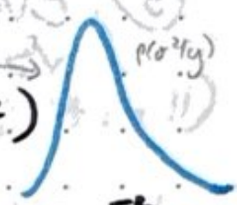


"marginal posterior"

* Marginal of σ^2:

$$P(\sigma^2|y) = \int P(\mu, \sigma^2|y)\, d\mu = \text{Inv-}\chi^2(\sigma^2 | n-1, s^2)$$

(inv. χ^2 distribution)



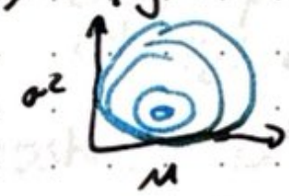
* Joint posterior

"conditional posterior" "marginal posterior"

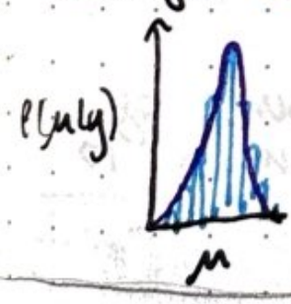
$$P(\mu, \sigma^2|y) = P(\mu|\sigma^2, y) P(\sigma^2|y) = N(\mu|\bar{y}, \sigma^2/n) \times \text{Inv-}\chi^2(\sigma^2|n-1, s^2)$$

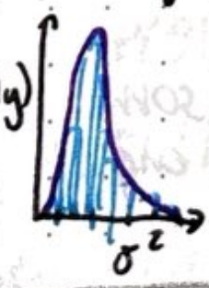
(from Bayes: $\propto P(y|\mu, \sigma^2)$)

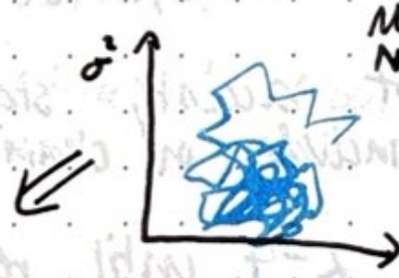
(from integral)



Metropolis 2D example
— analytic ○ posterior histogram







MC chain
$N_{mc} = 2000$
acc. ratio = 0.39

NOTE ON COV MATRICES (COVARIANCE OF MULTI DIM GAUSSIANS)

Reminder: if $\underline{x} \sim N(\underline{0}, \Sigma_x)$ then $\Sigma_x = \text{cov}(\underline{x}, \underline{x})$. (some prefer to denote this $\text{Cov}(\underline{x}, \underline{x}^T)$)

but I prefer to think "componentwise" i.e. $\Sigma_{ij} = \text{Cov}(x_i, x_j)$

e.g. $\begin{pmatrix} x \\ y \end{pmatrix} \sim N(\underline{0}, \Sigma_x)$ $\quad \Sigma_x = \begin{pmatrix} \text{cov}xx & \text{cov}xy \\ \text{cov}xy & \text{cov}yy \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \rho\sigma_y^2 \end{pmatrix}$

Think in components!

$$\text{Cov}[\theta_i^*, \theta_j^*] = \text{Cov}[\theta_i + L_{ik} z_k, \theta_j + L_{jl} z_l] = \text{Cov}[L_{ik} z_k, L_{jl} z_l] = L_{ik} L_{jl} \text{Cov}[z_k, z_l] = L_{ik} L_{jk}$$

$$= L_{ik} L^T_{kj}$$

$\text{cov}[ax, b+Y) = \text{cov}(X,Y)$ since θ_i const. (just a number)

$\text{cov}(ax, bY) = ab\,\text{cov}(X,Y)$

since z_i i.i.d $\text{cov}[z_i, z_j] = \delta_{ij}$

(Next algo in this case) algo generates a chain is "Markov process" + combine this w/ "monte carlo" nature of simulating new positions = gives the name MCMC ← circle note

METROPOLIS-HASTINGS ALGORITHM (slides)

More General Jumping rule $J(\theta^*|\theta_i)$

(Need not be symmetric $J(\theta_a|\theta_b) \neq J(\theta_b|\theta_a)$)

1. Choose random starting point θ_0
2. At step $i = 1, \ldots, N$ propose a new parameter value
$$\theta^* \sim J(\theta^*|\theta_{i-1})$$
3. Evaluate M-H ratio of posteriors at proposed vs current values.
$$r = \frac{P(\theta^*|y)/J(\theta^*|\theta_{i-1})}{P(\theta_{i-1}|y)/J(\theta_{i-1}|\theta^*)}$$
4. Accept θ^* with probability
$$\min(r,1): \theta_i = \theta^*$$
If not accept, stay at same value $\theta_i = \theta_{i-1}$ and include in chain.
5. Repeat 2-4 until reach some measure of convergence and gather enough samples to compute your inference.

METROPOLIS HASTINGS

Theory: see wikipedia

- want markov process (chain) to reach stationary distribution which is equal to the posterior
- a sufficient condition for stationary process is condition of detailed balance (i.e. prob being in state θ_i & transitioning to state θ_{i+1} is equal to prob being in state θ_{i+1} & transitioning to state θ_i). → the chain reversible
- process, 'flow of samples' between states as chain progresses is such that overall dist of samples is stationary (stationary dist)
- also want this stationary dist to be unique (i.e. no rising markov chain converging to dist: other than the posterior) → guaranteed by the ergodicity of the markov process (ergodicity = avg behavior observed over many diff instances ('ensembles') of the system. so avg behavior over time is the same as avg trajectory. ⊜ the same for many starting points → we always converge to same stationary dist so it is unique)

design a markov process (by constructing transition probabilities) that fulfills these 2 conditions and so stationary dist is the posterior → this is how metropolis hastings algo is

* d-dim Metropolis is just a special case, where $J(\theta^*|\theta_i) = N(\theta^*|\theta_i, \Sigma_p) = N(\theta_i|\theta^*, \Sigma_p) = J(\theta_i|\theta^*)$ is a symmetric proposal dist.

* More general asymmetric proposals, allow "biased" proposals → more probable to propose towards a certain direction

* With some knowledge of structure of the posterior, can sometimes engineer a clever proposal $J(\theta^*|\theta_i)$

(e.g.)

GIBBS SAMPLING

use when: { (joint) can't sample from $P(\theta, \phi)$ directly, but can sample from $P(\theta|\phi)$, $P(\phi|D)$ directly. (conditionals) }

* Special case of Metropolis-Hastings (Gelman BDA ch 11)

Multi-dim. sampling, when you can utilise the set of conditional posterior distributions as proposal dist.ns. for each parameter - Metropolis Hastings ratio = 1 (always accept)

* If joint posterior is $P(\theta, \phi|D)$

* And you can solve for tractable (you can draw) conditionals:

$$P(\theta|\phi, D)$$

$$P(\phi|\theta, D)$$

* Jump along each parameter-dimension one at a time

DERIVE: (algo. reaches dist π, if $\pi(\theta)P(\theta|\phi) = \pi(\phi)P(\phi|\theta)$ then starting dist $\pi(\theta)$ is $P(\theta|D)$) → to construct algo. choose starting dist to be posterior

$\frac{P(\theta_{i+1}|\theta_i)}{P(\theta_i|\theta_{i+1})} = \frac{P(\theta_{i+1})}{P(\theta_i)}$ so $\pi(\theta_i)P(\theta_{i+1}|\theta_i) = \pi(\theta_{i+1})P(\theta_i|\theta_{i+1})$

(1) detailed balance: $P(\theta_i)P(\theta_{i+1}|\theta_i) = P(\theta_{i+1})P(\theta_i|\theta_{i+1})$ so

(2) prob transition: $P(\theta_{i+1}|\theta_i) = J(\theta_{i+1}|\theta_i) A(\theta_{i+1}|\theta_i)$ ← prob propose, prob accept

(3) so $\frac{A(\theta_{i+1}|\theta_i)}{A(\theta_i|\theta_{i+1})} = \frac{P(\theta_{i+1})\,J(\theta_i|\theta_{i+1})}{P(\theta_i)\,J(\theta_{i+1}|\theta_i)}$ ☺

(4) choose acceptance ratio that satisfies this condition: Metropolis choice is $A(\theta_{i+1}|\theta_i) = \min\left(1, \frac{P(\theta_{i+1})\,J(\theta_i|\theta_{i+1})}{P(\theta_i)\,J(\theta_{i+1}|\theta_i)}\right)$

(works b/c either $A(\theta_{i+1}|\theta_i) = 1$ or $A(\theta_i|\theta_{i+1}) = 1$ so condition is always satisfied)

□

(slides)

saves computational resources → always accept

2-dim GIBBS SAMPLER

1. Choose random starting point (θ^0, ϕ^0) — don't overthink use this
2. At cycle t, update
$$\theta^t \sim P(\theta | \phi^{t-1}, D)$$
3. Then update
$$\phi^t \sim P(\phi | \theta^t, D)$$
(each complete set (pair) of updates is called a Gibbs cycle)
4. Record current values of chain (θ^t, ϕ^t)
5. Repeat 2-4 until reach some measure of convergence (G-R) and gather enough indep. samples to compute your inference (reduce Monte Carlo error)!

d-DIM GIBBS SAMPLER

Parameter vector $\theta = (\theta_1, \dots, \theta_d)$

Current state at j-th update within cycle t:

$(\theta_j^t, \underline{\theta}_{-j}^t)$

(θ_j^t: value of component j at iteration t; $\underline{\theta}_{-j}^t$: value of all components except j at iteration t (after j updates); d-1 dim vector)

$$\underline{\theta}_{-j}^t \equiv (\theta_1^{t+1}, \dots, \theta_{j-1}^{t+1}, \theta_{j+1}^t, \dots, \theta_d^t)$$

(iteration t+1 values for components already updated (?); iteration t (parent) values for components not yet updated (?))

1. Choose random starting point θ_0 $\underline{\theta}^0$ (?) no? only need θ_0
2. At cycle t, update through the d-parameters: For each $j = 1, \dots, d$ move the jth parameter to
$$\theta_j^{t+1} \sim p(\theta_j | \underline{\theta}_{-j}^t, D)$$
(update θ_j conditional on current values of all other parameters)
3. After updating all d parameters, record current state $\underline{\theta}^{t+1}$
4. Repeat 2-3 until reach convergence and enough samples

Gelman slightly diff notation:
$\theta_{-j}^{t-1} \equiv (\theta_1^t, \dots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \dots, \theta_d^{t-1})$

propose at iteration t: $\theta_j^t \sim p(\theta_j | \theta_{-j}^{t-1})$

(e.g. $d=3$) start $\theta^0 = (\theta_1^0, \theta_2^0, \theta_3^0)$

$\theta_1' \sim p(\theta_1 | \theta_2^0, \theta_3^0)$
$\theta_2' \sim p(\theta_2 | \theta_1', \theta_3^0)$
$\theta_3' \sim p(\theta_3 | \theta_1', \theta_2')$
} one cycle $t=1$

etc.

(slides)

Gibbs Sampling: Example

(Gelman BDA section 11.1)

consider single observation y_1, y_2

Likelihood: $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$

(ρ known)

Priors: $P(\theta_1) = P(\theta_2) \propto 1$

Posterior: $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \Big| y \sim N\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$

$$P(\underline{\theta}|y) = P(\theta_1|\theta_2, y)P(\theta_2|y) = P(\theta_2|\theta_1, y)P(\theta_1|y)$$

↳ since bivariate gaussian, joint can be decomposed into conditionals

(simple to draw directly from joint but for purpose of demonstrating gibbs sampling, find conditionals)

↳ Conditional Posteriors (eg. properties of MVGs): ← SEE MOODLE

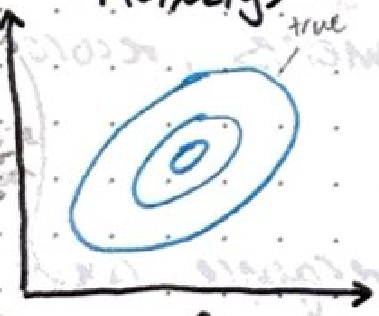
1D gaussians we can draw from →

$$\theta_1|\theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1-\rho^2)$$

$$\theta_2|\theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1-\rho^2)$$

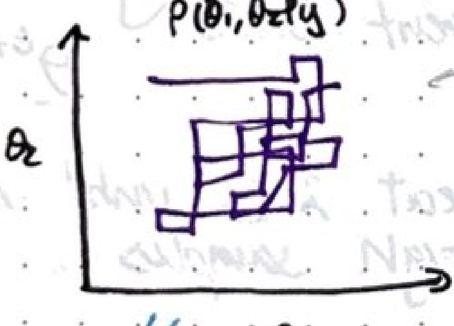
↳ improving MCMC ... how we could improve efficiency of sampler eg. sampler reparametrise / redefine rotation parameters

$P(\theta_1, \theta_2|y)$



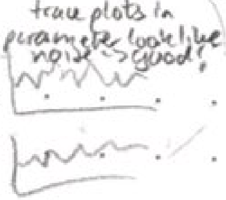
⟹

Gibbs sampling trace path:



← right angle! only update 1 param at a time

trace plots in each parameter look like noise → good!



Reminder: FLAW

if ρ large

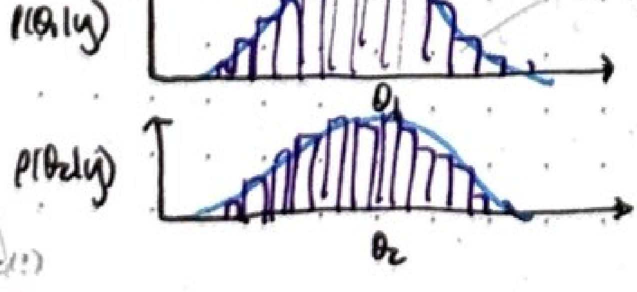
eg. $\rho = 0.99$

gibbs sampler → performs badly

using gibbs sampling: can only move on horizontal/vertical axis but dist. is diagonal so hard for gibbs sampler to explore the space!

↳ note: will also be a problem for metropolis hastings but especially a problem here!)

Marginal posterior densities



Lecture 18

5.3.25

Today: Continuing Bayesian Computation: Metropolis-Hastings, Gibbs Sampling, applications to SN cosmology, Assessing Convergence and Mixing.

GIBBS SAMPLING STEP AS A SPECIAL CASE OF METROPOLIS-HASTINGS (Gelman 11.3)

At the jth parameter update within iteration t (target $\pi(\vec{\theta}) = P(\vec{\theta}|D)$)

M-H ratio $$r = \frac{\pi(\vec{\theta}^*)/J(\vec{\theta}^*|\vec{\theta}^{t-1})}{\pi(\vec{\theta}^{t-1})/J(\vec{\theta}^{t-1}|\vec{\theta}^*)}$$

Recall $\vec{\theta}_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})$ — current components of $\vec{\theta}$ except θ_j

have updated up to component $j-1$

In Gibbs, we choose

only move to values in parameter space where all values are fixed?

$$J(\vec{\theta}^*|\vec{\theta}^{t-1}) = \begin{cases} \pi(\theta_j^*|\vec{\theta}_{-j}^{t-1}) & \text{if } \vec{\theta}_{-j}^* = \vec{\theta}_{-j}^{t-1} \\ 0 & \text{otherwise} \end{cases}$$

can only move to new proposals where ones we are not updating have to stay the same

more detail in notes on moodle

$$r = \frac{\pi(\vec{\theta}^*)/\pi(\theta_j^*|\theta_{-j}^{t-1})}{\pi(\vec{\theta}^{t-1})/\pi(\theta_j^{t-1}|\theta_{-j}^*)}$$

$\vec{\theta}_{-j}^* = \vec{\theta}_{-j}^{t-1}$ (use this to expand terms into marginals)

$$= \frac{\pi(\theta_j^*|\theta_{-j}^*)\pi(\theta_{-j}^*)/\pi(\theta_j^*|\theta_{-j}^{t-1})}{\pi(\theta_j^{t-1}|\theta_{-j}^{t-1})\pi(\theta_{-j}^{t-1})/\pi(\theta_j^{t-1}|\theta_{-j}^*)} = \frac{\pi(\vec{\theta}_{-j}^*)}{\pi(\vec{\theta}_{-j}^{t-1})} = 1$$

ALWAYS ACCEPT

saves on computation → always accept, don't need to calc. ratio

another variation: (slides)

Gibbs -> partitions params -> useful if natural partition of params in problem, but [illegible] if we can't solve for conditionals -> replace substep w/ metropolis rule.

Metropolis within Gibbs

$\theta = (\theta_1, \ldots, \theta_d)$ $\theta_{-j} \equiv (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d)$

* When you can't solve for tractable conditional distributions for all θ_j:

$$P(\theta_j | \theta_{-j}, D)$$

* Replace each substep for updating each jth parameter θ_j with a separate Metropolis rule, compute Metropolis ratio and accept/reject

* Cycle through all parameters and repeat for all N MCMC steps

multiple accept/rejects within full scan of parameters

PREVIOUS PAGE

Gibbs as special case of M-H:

def $\underline{\theta}^t_{-j}$ to mean this? (slightly diff to how we defined before? no?)

[$\underline{\theta}^t_{-j}$ is current components of $\underline{\theta}$ at cycle t except θ_j!]

MH: $r = \frac{P(\underline{\theta}^t)}{P(\underline{\theta}^{t-1})} \frac{J(\underline{\theta}^{t-1} | \underline{\theta}^t)}{J(\underline{\theta}^t | \underline{\theta}^{t-1})}$

(for Gibbs)
Propose proposal distribution $J(\underline{\theta}^t | \underline{\theta}^{t-1}) = \begin{cases} P(\theta_j^t | \underline{\theta}_{-j}^{t-1}) & ; \text{ if } \underline{\theta}_{-j}^t = \underline{\theta}_{-j}^{t-1} \\ 0 & ; \text{ otherwise} \end{cases}$ -> all components that are not j stay in their current state

-> i.e. only propose new points where all components are the same except θ_j, where $\underline{\theta}_{-j}^t = \underline{\theta}_{-j}^{t-1}$, θ_j is picked according to $P(\theta_j^t | \underline{\theta}_{-j}^{t-1})$

so $r = \frac{P(\theta_j^t, \underline{\theta}_{-j}^t)}{P(\theta_j^{t-1}, \underline{\theta}_{-j}^{t-1})} \frac{P(\theta_j^{t-1} | \underline{\theta}_{-j}^t)}{P(\theta_j^t | \underline{\theta}_{-j}^{t-1})}$

↑ components at t-1 ↑ components not j at t-1 } i.e. all components $\underline{\theta}^t$

$t=0$	$\theta_1^0 \; \theta_2^0 \; \theta_3^0$	($\underline{\theta}_j^0$)
$t=1$	$\theta_1^1 \; \theta_2^0 \; \theta_3^0$	<- ($\underline{\theta}_j^1$)

expand as conditionals:

$r = \frac{P(\theta_j^t | \underline{\theta}_{-j}^t) P(\underline{\theta}_{-j}^t) P(\theta_j^{t-1} | \underline{\theta}_{-j}^t)}{P(\theta_j^{t-1} | \underline{\theta}_{-j}^{t-1}) P(\underline{\theta}_{-j}^{t-1}) P(\theta_j^t | \underline{\theta}_{-j}^{t-1})} = 1$ => ALWAYS ACCEPT

(all cancel since $\underline{\theta}_{-j}^t = \underline{\theta}_{-j}^{t-1}$)

(everything we propose we accept so algo we saw before needed to make decision for how to propose, i.e. cycle through params! I think)

so could have defined θ_j^t same as before as a way to introduce this algo to our page!?

d-dim Metropolis-within-Gibbs sampler (slides)

$\theta = (\theta_1, \ldots, \theta_d)$ $\quad \theta_{-j}^t \equiv (\theta_1^{t+1}, \ldots, \theta_{j-1}^{t+1}, \theta_{j+1}^t, \ldots, \theta_d^t)$

1. Choose a random starting point θ_0
2. At cycle $t = 1, \ldots, N$, cycle through the d-parameters:
   - **A.** For each $j = 1, \ldots, d$, propose a new jth parameter value from a 1-dimensional Gaussian
     $$\theta_j^* \sim N(\theta_j^t, \tau_j^2)$$
     proposal scale τ^2: variance of gaussian
   - **B.** Evaluate ratio of posteriors at proposed vs current values:
     $$r = P(\theta_j^*, \underline{\theta}_{-j}^t | D) / P(\theta_j^t, \underline{\theta}_{-j}^t | D)$$
     $$= P(\theta_j^* | \underline{\theta}_{-j}^t, D) / P(\theta_j^t | \underline{\theta}_{-j}^t, D)$$
     i.e. $\frac{P(\text{new point})}{P(\text{old point})}$
     (note: only component j changes! $\underline{\theta}_{-j}$ stays the same)
   - **C.** Accept $\theta_j^{t+1} = \theta_j^*$ with prob $\min(r, 1)$,
     otherwise $\theta_j^{t+1} = \theta_j^t$
3. After full cycle, record current values θ^{t+1}
4. Repeat steps 2 for all parameters until convergence and enough samples

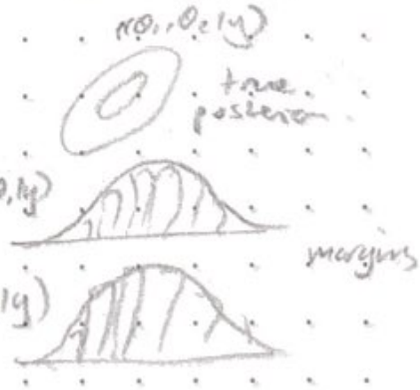
code demo example in lecture (same as GO!)

trace plot

Note: adjust τ^2:
empirically aim for 30% - 40% acceptance ratio

too small steps will accept every point but not explore param space, too long

too large never accept, never move anywhere



advantage of Gibbs: → always accept
→ fact you use conditional derived from posterior gives algorithm information about which direction is likely to have more probability

disadvantage: → requires analytic properties,
but if we can't derive them, can still use metropolis within Gibbs

* Mixed Gibbs Sampler:

→ Can replace sampling from conditionals with accept/reject for just the parameters with intractable conditionals

in cycle t
  for each j
    if we do have analytic conditional
      just propose new point from that & always accept

    if not
      propose new point from 1D normal
      and accept w/ prob min(r, 1)
      (aka metropolis)

$$r = \frac{P(\theta_j^{t+1}, \theta_{-j}^{t})}{P(\theta_j^{t}, \theta_{-j}^{t})}$$

Tuning d-dim Metropolis

* $\theta^* \sim N(\theta_i, \Sigma_p)$: If proposal scale Σ_p is too large, will get too many rejections and not go anywhere. If proposal scale too small, you will accept very many small moves: inefficient random walk.

* Laplace Approximation

$P(\theta|D) \approx N(\theta|\theta_{MAP}, A^{-1})$

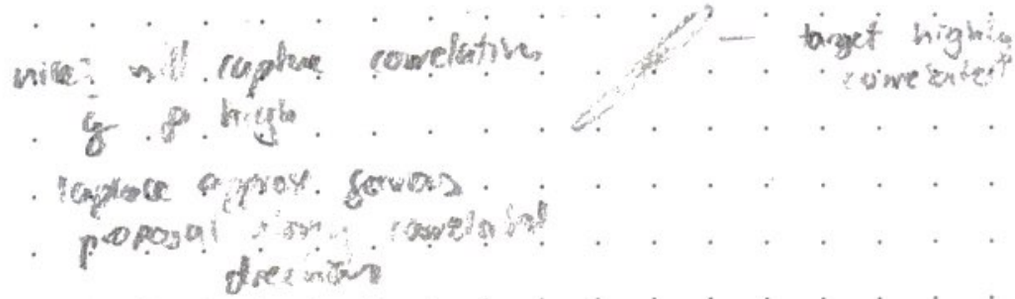
θ_{MAP} = posterior mode

$$A_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln P(\theta|D)\Big|_{\theta=\theta_{MAP}}$$

approx posterior cov matrix A^{-1}

* Choose $\Sigma_p = c^2 A^{-1}$ $\quad c \approx 2.4/\sqrt{\dim(\theta)}$

* Scale Proposal to aim for acceptance ratio of 44% in 1D, 23% in dim > 5.



LAPLACE APPROXIMATION

Unnormalised Posterior: $\hat{P}(\theta|D) = P(D|\theta)P(\theta)$

Find MAP estimate: "max. a posteriori"

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \ln \hat{P}(\theta)$$

← $\hat{P}(\theta|D)$ with Dobs plugged in?

Taylor expansion (first deriv. is zero):

$$\ln \hat{P}(\theta|D) \approx \ln \hat{P}(\theta_{MAP}|D) - \tfrac{1}{2}(\theta-\theta_{MAP})^T A (\theta-\theta_{MAP}) + \cdots$$

$$A_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln \tilde{P}(\theta|D)\Big|_{\theta=\theta_{MAP}}$$

$$\hat{P}(\theta|D) \approx \hat{P}(\theta_{MAP}) \times \exp\left(-\tfrac{1}{2}(\theta-\theta_{MAP})^T A (\theta-\theta_{MAP})\right)$$

Gaussian Approximation:

(approx normalised posterior)

$$P(\theta|D) \approx N(\theta|\theta_{MAP}, A^{-1})$$

A^{-1} is an approximate posterior covariance matrix.

reminder:
multi dim. taylor exp:
- let $\ell(\theta) = \ln \hat{P}(\theta|D)$
- $\ell(\theta) = \ell(\theta_{MAP}) + \nabla\ell(\theta_{MAP}) \cdot (\theta-\theta_{MAP}) + \tfrac{1}{2}(\theta-\theta_{MAP})^T H(\theta_{MAP})(\theta-\theta_{MAP}) + \cdots$

$\nabla\ell(\theta_{MAP})$: grad of ℓ evaluated at θ_{MAP} → 0

$H(\theta_{MAP})$: Hessian of ℓ (matrix of 2nd order derivatives) evaluated at θ_{MAP}

$\therefore \ell(\theta) \approx \ell(\theta_{MAP}) - \tfrac{1}{2}(\theta-\theta_{MAP})^T A (\theta-\theta_{MAP})$

A is -ve H eval at θ_{MAP}

Astrostatistics Case study: Supernova Cosmology

measure brightness of SN:

Cosmology: params control relationship between distance & redshift
dark energy EOS parameterised by w ($P = w\rho$)

modern cosmology → is $w = -1$?

type Ia SN almost standard candles
↳ measure their magnitude & plot vs redshift
curve changes depending on how we have
params Ω_m, Ω_Λ → so observed data constrains
the parameters → Bayesian analysis, told us $\Omega_\Lambda > 0$

comoving distance is integral found by from zero to redshift
distance probed by SN is "luminosity distance"
get d_L by multiplying that by dimensionless (ok with MLT)
distance redshift also influenced by curvature eg universe Ω_k

choose diff params plot
theoretical distance models
vs. redshift
diff curve for diff flatness etc.

data:

want to fit data to curve to infer params(?)

(slides) (lots of inconsistent notation here - ignore it)

For us: simplify, let's say SNae are standard candles.

$M_s \sim N(M_0, \sigma_{int}^2)$ — Population Distribution

$m_s = M_s + \mu(z_s; \theta)$ — (Log) inverse square law

(μ: distance modulus)

Assume data (apparent magnitudes and redshifts) $\{m_s, z_s\}$ are measured perfectly

$\theta = (H_0, \Omega_m, \Omega_\Lambda, w)$ — Cosmological Parameters

(save as Ω_Λ (in consistent notation nothing more))

First assume $w = -1$ (cosmological constant)
Derive model, likelihood, posterior

$\vec{\Omega} = (\Omega_m, \Omega_\Lambda, w)$, $h \equiv H_0/(100 km\,s^{-1}\,Mpc^{-1})$

$\mu(z_s; H_0, \vec{\Omega}) = 25 + 5\log_{10}\left(\frac{c}{H_0}\hat{d}_L(z_s; \vec{\Omega})\right)$

Mpc — dimensionless luminosity distance

$m_s = M_s + \mu(z_s; H_0, \vec{\Omega})$

$P(m_s | z_s) = \ldots$

$d_L = \frac{c}{H_0}\hat{d}_L$ (dimensionful)

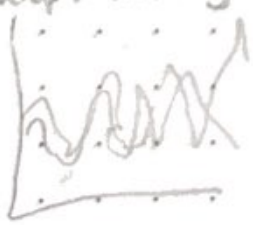
not true distance but distance inferred from brightness

luminosity distance = distance to an object w/ redshift z inferred from the standard candle method (need to take into account curvature (assume 1), scale factor (avg redshift), Ω_m, Ω_Λ, etc.)

M_0 and $\log_{10}(h)$ cannot be separately constrained! (not separately identifiable)

↓ 'degenerate'

multiple walkers shown
overlap: "mixing" → good sign, not getting stuck in parameter space



ex sheet 3

trace plot 4D metropolis
use tuplace · approx, time metropolis to sample posterior

(SHEET 3) see slides for plots of answers

SN cosmology reparametrisation

$M_s \sim N(M_0, \sigma_{int}^2)$

$M_s = M_0 + \varepsilon_{int}$

$m_s = M_0 + \mu(z_s | H_0, \vec{\Omega}) + \varepsilon_{int}$

$\vec{\Omega} = (\Omega_M, \Omega_\Lambda, w)$

$M_0 + \mu(z | H_0, \vec{\Omega}) = M_0 + 25 + 5\log_{10}\left(\frac{c}{100}\frac{100}{H_0}\tilde{d}_L(z_s, \vec{\Omega})\right)$

↖ 100 kms^{-1}Mpc^{-1}

$= \underbrace{M_0 - 5\log_{10}\left(\frac{H_0}{100}\right)}_{M_0 - 5\log_{10} h} + 25 + \underbrace{5\log_{10}\left(\frac{c}{100}\tilde{d}_L(z_s, \vec{\Omega})\right)}_{\delta(z_s, \vec{\Omega})}$

rename $\to M_0$

$m_s = M_s + 5\log_{10}\left(\frac{d}{10pc}\right)$

$= M_s + 5\log_{10}\left(\frac{d}{Mpc}\right) + 25$

$= M_s + 5\log_{10}\left(\frac{c/H_0 \tilde{d}_L}{Mpc}\right) + 25$

important?

Why: only have SN here (in sheet 1 had cepheids as well for which we knew their true distances) but have no cepheids to constrain abs mag so degeneracy between mean abs mag e.g. M_0 & H_0 cannot distinguish between them w/ SN data alone.

So can only constrain the combination

(a reparameterise problem, but we can only constrain the linear combination of mean SN abs mag M_0 & $\log_{10} H_0$

M_0 and $5\log_{10} h$ are degenerate!

board

$P(m_s | z_s; M_0, \vec{\Omega}, \sigma_{int}^2) = N(m_s | M_0 + \delta(z_s; \vec{\Omega}), \sigma_{int}^2)$

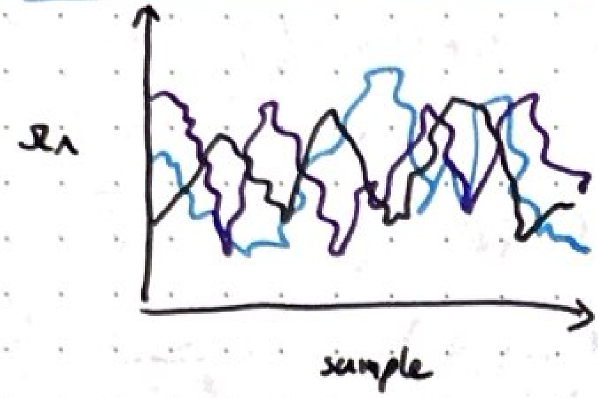
Likelihood: $L(M_0, \vec{\Omega}, \sigma_{int}^2) = \prod_{s=1}^{N} P(m_s | z_s; M_0, \vec{\Omega}, \sigma_{int}^2)$

(Indep.) priors: $M_0, \Omega_\Lambda \sim U(-\infty, \infty)$, $\Omega_M, \sigma_{int}^2 \sim U(0, \infty)$ non informative priors

Posterior: $P(M_0, \vec{\Omega}, \sigma_{int}^2 | \{m_s, z_s\}) \propto L(M_0, \vec{\Omega}, \sigma_{int}^2) \times P(M_0) P(\Omega_M) P(\Omega_\Lambda)$

(slide)

Multiple Indep. Chains



(in param space)

chains overlap: "mixing" = good!

Assessing Convergence w/ multiple chains: Gelman-Rubin (G-R) ratio

Suppose we have simulated m parallel sequences, each of length n (after discarding the first half of the simulations). For each scalar estimated ψ, we label the simulation draws as ψ_{ij} ($i = 1, \dots, n$; $j = 1, \dots, m$), and we compute B and W, the between- and within-sequence variances:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{.j} - \bar{\psi}_{..})^2$$

(i.e. variance of the sum of all samples in each chain)

$$\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij} \quad \text{(avg over samples for given chain } j\text{)}, \qquad \bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j} \quad \text{(avg over everything (all samples in all chains))}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2$$

(variance of each chain averaged over all chains)

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{.j})^2$$

We can estimate $\text{var}(\psi|y)$, the marginal posterior variance of the ~~estimated~~ estimand, by a weighted average of W and B, namely

$$\widehat{\text{var}}^+(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B$$

("estimate variance of underlying distribution from which samples are served") (chatgpt) total variance, accounts for between & within-chain variance

G-R ratio: $\hat{R} = \sqrt{\frac{\widehat{\text{var}}^+(\psi|y)}{W}} \approx 1$

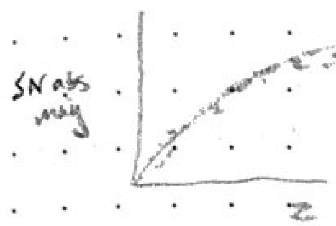
if $W \approx B$ then say it has converged / G-R ratio ≈ 1; if not maybe run chains for longer

not good prog / but guaranteed good chain if $\hat{R} \approx 1$

compare variance within chain & variance between chains to give measure of convergence / how well they have mixed

- tune proposal scale to give acceptance ratio ~30%
- run indep chains for parameters → want GR ratio ≈ 1
- cut off ~20% for burn in (rule of thumb)

see slides / sheet 3



nobel prize 2011

MCMC → can also find posteriors on derived quantities (useful)
see slides / lecture → ex sheet 3

(19) ↳ can compute posterior of any derived quantity e.g. deceleration parameter
find high prob $q_0 < 0$ → universe is decelerating

(for each chain) (sum of all samples in that chain) (avg of sums of samples in each chain over all chains)

$$\Longrightarrow B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\psi_{ij} - \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\psi_{ij}\right)^2$$

think of B as variance in the mean of each chain times n since $Var(\bar{x}) = \frac{1}{n}Var(x)$
(chain mean is calculated from n samples)
this restores the scale to match the variance of individual samples
(chat GPT)

summary: (W: variance within chain
B: variance between chains scaled to individual sample variance)

(not sure if this is correct? chat gpt)

$$Var(x) = E[Var(x|chain)] + Var(E[x|chain])$$ ← LAW OF TOTAL VARIANCE (/ variance decomposition)

↑ expected within-chain variance $\approx \frac{n+1}{n} W$
↑ correction factor since W based on sample variances?

↑ variance of chain means $\approx B/n$

Lecture 19

7.3.25

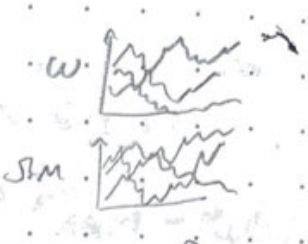
Today: continue MCMC – applications to SN cosmology, assessing convergence & mixing, comparison of algos

(sheet 3)

New case:

Now assume flat univ. $\Omega_\Lambda = 1 - \Omega_M$

but unknown $M_0, \sigma_{int}^2, \Omega_M, w$



changing $w \rightarrow$ v. small effect, so hard to constrain w

now instead use metropolis-within-gibbs 4x1D for each parameter (fix all other params to their current value and update one)

tune 4x proposal scales & 4x acceptance ratios

(visually → doesn't look like the best sampling on slides (not mixed well))

2D plot → notice banded structure in samples: signature of the algorithm (due to rejection)

can use estimated cov. matrix (from initial MCMC or Laplace approx) for correlated proposal dist. in 4D metropolis!

(helps if know what direction to sample in)

use this as proposal for 4D metropolis → see much better movement/mixing, chains look more similar → improvement!

estimate cov matrix from samples from this initial run & use as a proposal dist. for new 4D metropolis run.

better because we have tuned proposal/jump dist to better match the true dist. so it guides us to sample in the right direction (effectively explore sample space)

(slides)

How Many Iterations to get an independent sample? Autocorrelation Function

For each scalar parameter θ

Chain: $(\theta_1, \dots, \theta_N)$

Consider sample mean $\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N} \theta_i$

If chain were serially uncorrelated: $\mathrm{Cov}(\theta_i, \theta_j) = 0$, $(i \neq j)$

then $\mathrm{Var}[\bar{\theta}] = \mathrm{Var}(\theta)/N$.

However, chain is typically serially correlated:

$\mathrm{Cov}[\theta_i, \theta_j] = C_{|i-j|} = \mathrm{Var}(\theta)\rho_{|i-j|}$.

Then for large N: $\mathrm{Var}[\bar{\theta}] = \mathrm{Var}(\theta) \times \tau/N = \mathrm{Var}(\theta)/N_{eff}$

where $\tau = 1 + 2\sum_{t=1}^{\infty} \rho_t$ is the AUTOCORRELATION TIMESCALE

basically tells us how many more steps do you need to take before your next indep. sample?

and $N_{eff} = N/\tau$ is the EFFECTIVE SAMPLE SIZE (ESS) (equivalent number of independent samples).

- large corr ρ_t gives small ESS: correlated samples provide less info than uncorrelated samples from underlying dist

(reminder $\mathrm{cov}(x,y) = \sqrt{\mathrm{var}x}\sqrt{\mathrm{var}y}\,\rho$)

$\mathrm{Cov}[\theta_i, \theta_j] = \mathrm{Var}(\theta)\rho_{|i-j|}$.

serially correlated: errors at one point in chain are related to errors at another point in chain. each value in chain depends on values of other values in chain as a fn of distance between those points in chain!

PROOF

$\Rightarrow$ Find $\mathrm{Var}[\bar{\theta}]$: reminder: variance of sum is sum of covariances!

$$\mathrm{Var}[\theta] = \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\theta_i\right] = \frac{1}{N^2}\mathrm{Var}\left[\sum_{i=1}^{N}\theta_i\right] = \frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}(\theta_i,\theta_j)\right) \quad \text{- serially correlated}$$

$$= \frac{1}{N^2}\left(\sum_{i\neq j}\mathrm{cov}(\theta_i,\theta_j) + \sum_{i=1}^{N}\mathrm{Var}(\theta_i)\right) = \frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{j\neq i}^{N}\mathrm{Var}(\theta)\rho_{|i-j|} + N\,\mathrm{Var}(\theta)\right)$$

$$= \frac{1}{N}\mathrm{Var}(\theta)\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\rho_{|i-j|} + 1\right]$$

eval this sum: $\rho_{|i-j|}$ is just a fn of distance. For infinitely long chain imagine going in each direction & adding up all distances: get factor of 2 going left vs right & factor of N since do this for each of N values in chain. (which is apparently ergodic. ok.)

for each point i in chain

so for large N: $\mathrm{Var}(\bar{\theta}) = \frac{1}{N}\mathrm{Var}(\theta)\left[2\sum_{t=1}^{\infty}\rho_t + 1\right]$ //

do this in practice:

Estimating the Autocorrelation / ESS

For each scalar parameter θ

Sample covariance of lag t:

$$\hat{C}_t = \frac{1}{N-t-1} \sum_{i=1}^{N-t} (\theta_i - \bar{\theta})(\theta_{i+t} - \bar{\theta})$$

↳ take all pairs that are distance t apart and estimate covariance

$\hat{C}_0$ = sample variance of θ

$\hat{\rho}_t = \frac{\hat{C}_t}{\hat{C}_0}$ Sample autocorrelation of lag t ("distance")

$\hat{\tau} = 1 + 2\sum_{t=1}^{T} \hat{\rho}_t$ Estimated autocorrelation time

Truncate at T lags s.t. $\hat{\rho}_T \approx 0.1$

$N_{eff} = N/\hat{\tau}$ Effective number of independent samples

Slowest parameter is the limiting one! (convergence time limited by slowest mixing param)

In practice, what matters is:

Compute time / Effectively Independent Sample

- want to estimate: $N_{eff} = N/\tau$
- need ρ_t, ($\tau = 1 + 2\sum_t \rho_t$)
  $C_{|i-j|}$ → use:
- $Cov(\theta_i, \theta_j) = Var(\theta)\rho_{|i-j|}$
- need to know no of samples that contribute to sum at each t:
  $1 \ldots N$
  → only one pair of samples distance $t = N-1$ apart ($\leq N-t$ pairs).
  → $N-2$ pairs of samples distance $t=1$ apart ($\leq N-t$ pairs)

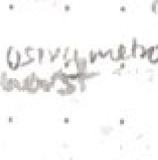
see slides

weigh up how fast algo is vs how many effective indep samples i.e. time it takes to get 1 indep sample! good to compare metric

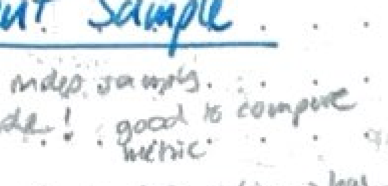
see slides → using metropolis within Gibbs, worst autocorr is ω → compare to better sampling algo → has larger effective sample number (faster rate of ...)

banded

better sampling → don't see banded structure in sampling.



try a diff sampling method
↳ 4D metropolis
↳ ~14x faster than MWG!
better looking graph, faster rate of indep samples (larger N_{eff})



still large ρ at distance 1000 → not good

for $N = 10{,}000$,
→ $N_{eff} \approx 13$
not good!
⟹ 43s to get one independent sample

can mix & match these algos: e.g.

Mixed Samplers

Target Posterior (intractable - cannot sample) $\pi(\alpha, \beta, \theta)$

Suppose these $\pi(\alpha|\beta,\theta)$, $\pi(\beta|\alpha,\theta)$, $\pi(\alpha,\beta|\theta)$ are tractable

but $\pi(\theta|\alpha,\beta)$ is not tractable

Mixed sampler:

1) Sample $\alpha^t \sim \pi(\alpha|\beta^{t-1}, \theta^{t-1})$
2) Sample $\beta^t \sim \pi(\beta|\alpha^t, \theta^{t-1})$
3) Metropolis (proposal/acc./rej.) update θ^t

Parameter Blocking

Target posterior (intractable - cannot sample): $\pi(\alpha, \beta, \theta)$

As before suppose $\pi(\theta|\alpha,\beta)$ not tractable.

Mixed Sampler with Blocking: (more efficient)

1) Jointly sample $\alpha^t, \beta^t \sim \pi(\alpha,\beta|\theta^{t-1})$
2) Metropolis (proposal/acc./rej.) update

have shown how MCMC works → now show why it works

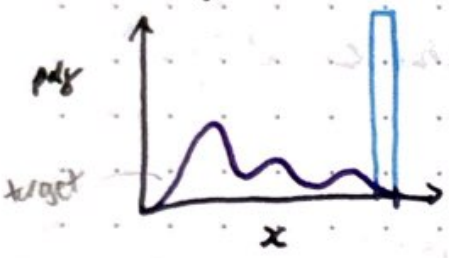
(step 1)

Run 10,000 chains:

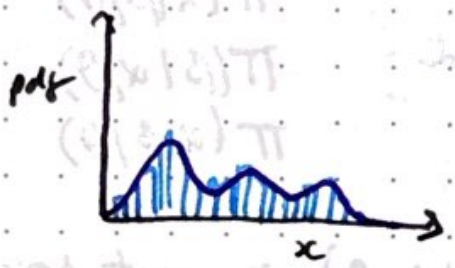
(start all chains at some value)

histograms of ensemble of chains at time t (not single chain as we are used to)

Histogram of initial values of chains (t=1)



Histogram of chain values at step t=100



looks like time-dist for 1 chain

(also looks like true dist.)

ERGODICITY:

Convergent distribution of chain ensemble at single time-slice = distribution-over-time of single chain = target (posterior) distribution (convergence to a stationary distribution)

Time average = Ensemble average

avg behavior over time is the same as behavior observed over many diff instances ("ensembles") of the system

big q why we do MCMC → want as long as we can our sample to match the distribution, but we don't want to run loads of chains 100 times to reach that dist. → we just want to run one chain & have it approximate the dist at large times

ergodicity ⇒ unique stationary distribution? (yes, correct)

A SKETCH OF MCMC THEORY

Markov chain: A sequence of random variables:

$\theta^0, \theta^1, \ldots, \theta^{t-1}, \theta^t \in \Theta$ (state space = parameter space)

values in chain are in state space

with Markov property: $P(\theta^t | \theta^{t-1}, \ldots, \theta^1, \theta^0) = P(\theta^t | \theta^{t-1})$

only depends on previous state, not all steps in between

Joint distribution: $P(\theta^t, \theta^{t-1}, \ldots, \theta^1, \theta^0)$

$= P(\theta^t | \theta^{t-1}) \times P(\theta^{t-1} | \theta^{t-2}) \times \ldots \times P(\theta^1 | \theta^0) P(\theta^0)$

↓ call $\pi_0(\theta^0)$ (pdf of initial state)

$= \pi_0(\theta^0) T_1(\theta^0 \to \theta^1) \times \ldots \times T_{t-1}(\theta^{t-2} \to \theta^{t-1}) \times T_t(\theta^{t-1} \to \theta^t)$

$T_t(\theta^{t-1} \to \theta^t) \equiv P(\theta^t | \theta^{t-1})$

← transition probabilities →

If $T_t(\theta^{t-1} \to \theta^t) = T(\theta^{t-1} \to \theta^t)$ not explicitly time-dependent → "time homogenous".

(e.g. prob transition state θ^0 to θ^1 is same whether at beginning or end of chain)

Marginal distribution:

$$P(\theta^t) = \int \pi_0(\theta^0) T(\theta^0 \to \theta^1) \times \ldots \times T(\theta^{t-1} \to \theta^t) \, d\theta^0 \ldots d\theta^t$$

Goal of MCMC: Want to engineer Markov chain s.t.

1. Limiting distribution $\pi_{lim}(\theta) = \lim_{t \to \infty} P(\theta^t = \theta)$ = Stationary distribution aka invariant distribution $\pi_{stat}(\theta)$

   (stationary distribution exists (& is unique))

2. $\pi_{lim}(\theta) = \pi_{stat}(\theta)$ is equal to target posterior $\pi_{post}(\theta)$

   (and it is equal to the target)

① system converges to unique stationary dist (?)
② that dist is the target

Good Properties for MCMC:

(local)

i.e. can explore param space in good time

1) IRREDUCIBLE

MC can get to any state from any other state in a finite # of steps.

2) APERIODIC

(def in contrast to what periodic is)

system does not return to the same state at fixed intervals

[Periodic: MC has positive probability of returning to a state only with period $k>1$]

period $k=1$

3) NOT TRANSIENT

def in contrast to transient

(aka recurrent → prob system will return to previous state = 1; weaker condition than pos. recurrent (includes infinite return time) (will return to state but might take infinitely long to get there)

[Transient: Non-zero probability of chain never returning to a particular state]

Wikipedia M-H algo:

markov process has a unique stationary dist provided

① stationary dist exists (this is true b/c of condition of detailed balance) (implied by)

② stationary dist is unique guaranteed by ergodicity which implies:

→ aperiodic

→ positive recurrent

(also need irreducible? to converge?) & stationary dist to positive?

demo MCMC codes on moodle

Lecture 20

10.3.25

"detailed balance": e.g. if oxygen molecule moves to other side of room another molecule will take its place → conc. of oxygen will stay the same, same ident. dist

histogram of MCMC samples will move around as chains sample diff. peaks but settle overall

Today: finishing Bayesian computation (MCMC), soon: gaussian processes

A sketch of MCMC theory cont.

(board)

(First, some jargon:)

Positive Recurrent: has finite expected time to return to each state

Ergodic: Aperiodic & positive recurrent (time avg. = expectation over ensemble)

If we have:

Markov chain which is / Markov chain

1) Irreducible
2) Aperiodic
3) Not transient

} MC converges to a unique, stationary distribution π_{stat} as the limiting dist. π_{lim} $(t \to \infty)$

positive recurrent (stronger condition) → not transient (alone does not guarantee convergence)

- We want to construct transition probabilities s.t. $\pi_{stat} = \pi_{post}$ (target) (want end state to be true posterior)
- To do that, show that target distribution is invariant under transition probabilities, i.e. target is the unique stationary distribution.

Stationary Distribution

(suppose we start at point a — what is prob like next pt is b? — but avg over all starting points a.)

Suppose $\theta^{t-1} = a \sim \pi(\theta)$ $\int P(\theta^{t-1}=a)\, T(\theta^t = b \mid \theta^{t-1}=a)\, d\theta^{t-1}$

Next step: $P(\theta^t = b) = \int P(\theta^{t-1})\, T(\theta^{t-1} = a \rightarrow \theta^t = b)\, d\theta^{t-1}$

(prob next point is b regardless of starting point)

$$= \int \pi(a)\, T(a \rightarrow b)\, da$$

iff $P(\theta^t = b) = \pi(b)$ then π is the stationary distribution π_{stat}.

→ (target dist π is invariant under transition probabilities)

(transition probabilities do not induce any net change in pdf over all states)

Detailed Balance:

A sufficient condition (for stationary dist to exist) - chain respects D.B. aka reversible, with π as stationary distribution

Reversible: $P(\theta^{t-1} = a, \theta^t = b) = P(\theta^{t-1} = b, \theta^t = a)$

|| same thing (i.e. ∃ stationary dist π that satisfies D.B. then process is reversible)

Condition of D.B.: (prob it goes one way = prob it goes the other way)

$$\pi(a)\, T(a \rightarrow b) = \pi(b)\, T(b \rightarrow a)$$

$\pi(a)T(b|a) = \pi(b)T(a|b)$ ← alternative notation

(flow of probability $a \rightarrow b$ perfectly balanced by flow from $b \rightarrow a$)

If $\theta^{t-1} = a \sim \pi$, $P(\theta^t = b) = \int \pi(a)\, T(a \rightarrow b)\, da$
$$= \int \pi(b)\, T(b \rightarrow a)\, da$$

$P(\theta^t = b) = \pi(b) \int P(a|b)\, da = \pi(b)$

→ π is the stationary distribution π_{stat}!

(board) — if transition prob & dist π satisfy detailed balance then under conditions 1,2,3 it will converge to stationary dist. Now we want to show under metropolis rule satisfies detailed balance

(SEE APPENDIX)

Metropolis algorithm respects D.B. w/ $\pi_{stat} = \pi_{post}$

(→ shown that if transition prob satisfies DB then stationary dist exists; → now show that our algo is constructed s.t. transition probabilities satisfy DB as the stationary dist)

Suppose wlog we pick points $a, b \in \Theta$ s.t. $\pi_{post}(a) < \pi_{post}(b)$, and $a, b \sim \pi_{post}(\Theta)$, and proposal distribution $J(b|a) = J(a|b)$ (symmetric). $(r>1)$.

(whole expression is symmetric in a,b; even though we narrow to particular case, this is true in general)

Forward:

$$P(\theta^{t-1}=a, \theta^{t}=b) = \pi_{post}(a)\,T(a\to b) = \pi_{post}(a)\,J(b|a)\times 1$$

(prob make transition; prob propose transition; prob accept transition)

$1 = \min\left(\frac{\pi_{post}(b)}{\pi_{post}(a)}, 1\right)$

Backward:

$$\begin{aligned} P(\theta^{t-1}=b, \theta^{t}=a) &= \pi_{post}(b)\,T(b\to a) \\ &= \pi_{post}(b)\times J(a|b)\times \frac{\pi_{post}(a)}{\pi_{post}(b)} \\ &= J(a|b)\,\pi_{post}(a) = J(b|a)\,\pi_{post}(a) \end{aligned}$$

Forward = Backward → transition probabilities respect D.B. w/ $\pi_{stat} = \pi_{post}$

(reminder: Josh Spiegel paper about limitations MCMC is useful to read!)

(Exercise:) (see appendix!)

1. Show DB also holds for Metropolis Hastings w/ asymmetric proposal $J(a|b) \neq J(b|a)$

   (consider full cycle of Gibbs, show $\pi_{post}(\vec{\theta})$ is $\pi_{stat}(\vec{\theta})$)

2. Gibbs sampling: suppose $\pi_{post}(\theta, \phi)$ is the posterior over two parameters θ, ϕ. the conditionals $\pi(\theta|\phi)$ and $\pi(\phi|\theta)$ are tractable. By considering a full cycle of the Gibbs sampler, show that $\pi_{post}(\theta, \phi)$ is the stationary dist.

show that one step of a Gibbs sampler is equivalent to a M-H step where the acceptance ratio is always 1.

(slides)

MCMC in Practice

1. Find the mode(s) using optimisation (peaks), can use Laplace approx. to obtain a proposal cov. matrix → for laplace approx need expression for posterior & can differentiate (not necessarily normalised?) $\underline{A} = -\underline{H}|_{\theta_{MAP}}$

   alternative to laplace if not tractable posterior and posterior looks very badly gaussian is to do initial run & estimate cov matrix from samples as proposal dist for new run!

2. Begin multiple (4-8 parallel) chains at starting positions dispersed around the mode(s) (can draw from approx. cov. matrix) ← from laplace ← or initial MCMC run

   can see if particular chain gets stuck

3. Scale Metropolis proposals to tune 25-50% acceptance rate (depending on dimensionality of jump).

4. Use proposal cov. matrix that reflects the shape of the posterior (e.g. banana posterior example L18/19)

5. After run, look at chains (if possible) to check for obvious problems — should look like noise

6. Compute Gelman-Rubin ratio comparing within-chain-variance to between-chain-variance to check that chains are well-mixed (should be very close to 1), and assess burn-in

   not good proxy → won't say for all if mixed but will say is not mixed.

7. Compute autocorrelation timescale and effective sample size to make sure you have enough independent samples for inference

8. If all checks out, remove burn-in, thin, and combine chains to compute posterior inferences.

Overview of MCMC

* MCMC Algorithms — most other algos are just special cases of M-H
  - Metropolis / M-H algorithms
  - Gibbs Sampling
  - Metropolis-within-Gibbs

  didn't cover: modern MCMC use gradient (autodiff) - Hybrid / HMC / NUTS

* Assessing / Comparing performance of MCMC algo.s
  - Gelman-Rubin statistics for comparing mixing of multiple chains
  - Autocorrelation time — how long to get an independent sample?
  - Effective sample size - how many indep. samples do I have?

* Detailed balance & theoretical considerations

end of MCMC!

(s next time: gaussian processes

Lecture 21

12.3.25

GAUSSIAN PROCESSES (slides)

What is a Gaussian Process?

* A G.P. is a (possibly infinite) collection of R.V.s $\{f_t\}$, (typically indexed w/ some ordering in time, space or wavelength), such that any finite subset of R.V.s have a jointly multivariate Gaussian dist.
* Any vector $f = \{f_t : t = 1, \ldots, N\}$ of a finite subset is multivariate Gaussian, therefore it is completely described by a mean $E[f]$ and cov. matrix $\text{Var}[f] = \text{Cov}[f, f^T]$.
* Elements of the cov. matrix are determined by a function of the coordinates e.g. $\text{Cov}[f_t, f_{t'}] = k(t, t')$ called the covariance function or kernel.
* A G.P. w/ mean function $m(t)$ and kernel $k(t, t')$ is denoted

$$f(t) \sim GP(m(t), k(t, t'))$$

* A G.P. provides a distribution over functions

the dist of a GP is the joint of all these (infinitely many) variables, i.e. it is a dist over fns with a continuous domain (i.e. time or space)

e.g. $f = \{f_1, f_2\}$

$$\begin{bmatrix} f(1) \\ f(2) \end{bmatrix} \sim N\left(\begin{bmatrix} m(1) \\ m(2) \end{bmatrix}, \begin{bmatrix} k(1,1) & k(1,2) \\ k(2,1) & k(2,2) \end{bmatrix} \right)$$

what is ?

since gaussian, GP is completely defined by its mean $m(t)$ and covariance $k(t, t')$

see notes on moodle (slides)

Review: Properties of multivariate Gaussians

Full prob. density: (Σ positive definite)

$$N(\underline{y}|\underline{\mu},\Sigma) \equiv [\det(2\pi\Sigma)]^{-1/2}\exp\left(-\tfrac{1}{2}(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right)$$

Joint dist. of components:

$$\underline{y} = \begin{pmatrix}\underline{U}\\ \underline{V}\end{pmatrix} \sim N\left(\begin{bmatrix}\underline{U}_0\\ \underline{V}_0\end{bmatrix}, \begin{bmatrix}\Sigma_U & \Sigma_{UV}\\ \Sigma_{VU} & \Sigma_V\end{bmatrix}\right)$$

If you observe/know/condition on $\underline{V}$:

Conditional dist. $\quad \underline{U}|\underline{V} \sim N(E[\underline{U}|\underline{V}], \mathrm{Var}[\underline{U}|\underline{V}])$

Conditional mean $\quad E[\underline{U}|\underline{V}] = \underline{U}_0 + \Sigma_{UV}\Sigma_V^{-1}(\underline{V}-\underline{V}_0)$

Conditional variance $\quad \mathrm{Var}[\underline{U}|\underline{V}] = \Sigma_U - \Sigma_{UV}\Sigma_V^{-1}\Sigma_{VU}$

If $\underline{V}$ = observed data, $\underline{U}$ = unobserved parameters, then $P(\underline{U}|\underline{V})$ is a posterior pdf!

reminder, marginals: $P(\underline{V}) = \int P(\underline{U},\underline{V})\,d\underline{U} = N(\underline{V}|\underline{\mu}_V, \underline{\Sigma}_V)$

$P(\underline{U}) = N(\underline{U}|\underline{\mu}_U, \underline{\Sigma}_U)$

(simply just drop irrelevant variable from mean vector & cov matrix! e.g. $\underline{\mu} = \begin{bmatrix}\underline{\mu}_U\\ \underline{\mu}_V\end{bmatrix} \to \underline{\mu}_U$

$\underline{\Sigma} = \begin{bmatrix}\underline{\Sigma}_U & \underline{\Sigma}_{UV}\\ \underline{\Sigma}_{VU} & \underline{\Sigma}_V\end{bmatrix} \to \underline{\Sigma}_U$)

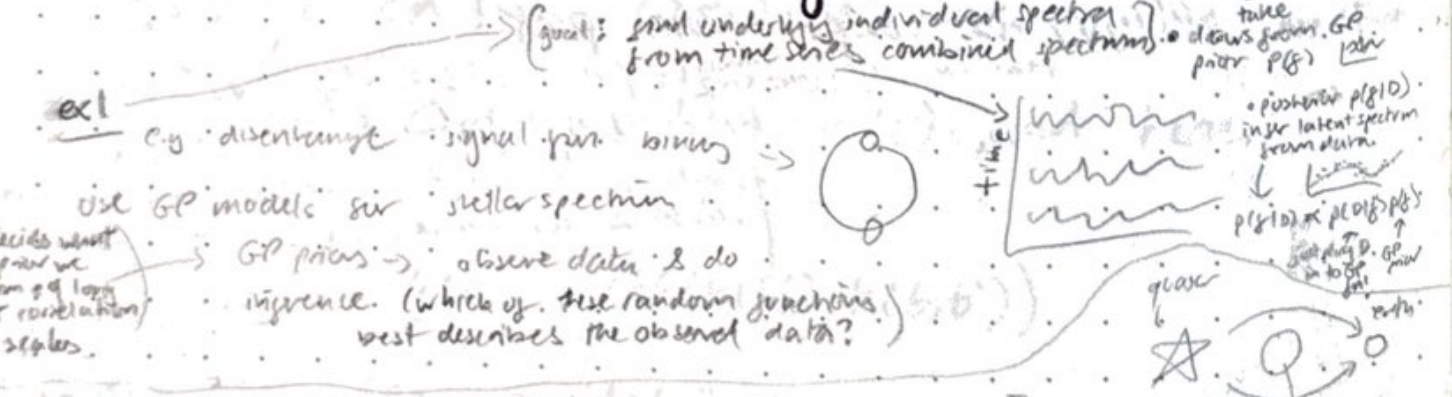
then conditionals same form. $p(\underline{U}|\underline{V}) = \dfrac{P(\underline{U},\underline{V})}{P(\underline{V})}$

(don't know $\underline{y}$, i need to know prior for this?)

What are GPs used for in Astrophysics?

(slides)

- Some physical models are very nearly gaussian e.g. CMB is G.P. on the sphere) or are gaussian (e.g. damped random walks as quasar light curves are approx gaussian)
- "Nonparametric" models: flexible fns to use when an accurate parametric astrophysically-motivated fn. is not available or is impossible.
- Nonparametric ⇒ no. of parameters / latent variables grows with the dataset.
- Interpolation / Emulation: to generate a smooth curve going through some observation or simulation points
- Correlated noise/error model: When you marginalise out the latent function, you are effectively accounting for noise/fluctuations correlated over time/space/wavelength → "nuisance function"



eg fit GP to SN light curve time series

G.P. as a prior on functions (slides)

But we only ever evaluate the fn. on a finite set of points!

A grid of times: $t = (t^1, \ldots, t^i, \ldots, t^N)^T$

A vector of fn. values evaluated on the time grid:

$$f = (f(t^1), \ldots, f(t^i), \ldots, f(t^N))^T$$

Assume a squared exponential kernel / cov. fn.:

$$\mathrm{Cov}[f(t), f(t')] = k(t,t') = A^2 \exp(-|t-t'|^2/\tau^2)$$

Assume a const. prior mean fn.:

$$\mathbb{E}[f(t)] = m(t) = c \quad (\text{often assume zero-mean } c=0)$$

Prior on function: $P(f|A,\tau) = N(f|1c, K)$

Drawing from prior: $f|A,\tau \sim N(1c, K)$

Cov. matrix K is populated by evaluating the kernel:

$$\mathrm{Cov}[f(t), f(t')] = k(t,t') = A^2 \exp(-|t-t'|^2/\tau^2)$$

For all pairs of points in t:

$$K_{ij} = k(t_i, t_j) = A^2 \exp(-|t_i - t_j|^2/\tau^2)$$

non-covariant kernel, set A, τ to particular values & generate say 20 random fns drawn GP prior →

(notice how fns change if we change τ, A!)

hyperparameters of kernel have significant influence on fits!

(slides)

Fitting a GP to data

① **PREDICTION**
If we knew the characteristic scales of the kernel (hyperparameters) (A, τ^2), then how do we fit the data at observed times to find the curve for unobserved times?

② **MODEL SELECTION**
Given the observed data, how do we fit for the characteristic scales of the kernel (hyperparameters)?

Posterior inference w/ G.P.s, Estimating the underlying curve:

f_0 = fn. at observed times t_0 (training set)

f_* = fn. at unobserved times t_* (prediction set)

Joint: $$\begin{pmatrix} f_0 \\ f_* \end{pmatrix} \sim N\left(\begin{bmatrix} 1c \\ 1c \end{bmatrix}, \begin{bmatrix} K(t_0,t_0) & K(t_0,t_*) \\ K(t_*,t_0) & K(t_*,t_*) \end{bmatrix} \right)$$

Populating the cov. matrix:

$K(t,t')$ has i,j-th entry $= k(t_i, t_j')$

Using the assumed kernel fn.

$$\mathrm{Cov}[f(t), f(t')] = k(t,t') = A^2 \exp(-|t-t'|^2/\tau^2)$$

now use properties of MVG. vectors
↓
partition

Posterior (conditional on observations, is also Gaussian):

$$f_* | f_0 \sim N(\mathbb{E}[f_* | f_0], \mathrm{Var}[f_* | f_0])$$

Posterior predictive mean:

$$\mathbb{E}[f_* | f_0] = 1c + K(t_*, t_0) K(t_0, t_0)^{-1} (f_0 - 1c)$$

Posterior predictive variance / covariance:

$$\mathrm{Var}[f_* | f_0] = K(t_*, t_*) - K(t_*, t_0) K(t_0, t_0)^{-1} K(t_0, t_*)$$

ignore measurement error -> bad
GP tries to fit by going through centre of every single data point

including measurement error: good

how do we account for this?

Accounting for measurement error

"Think of measurement error as adding gaussian process to gaussian process"

$$y_0 | f_0 \sim N(f_0, W)$$

y_0 are measured values of f_0 at times t_0

at each observation $y_0^i | f_0^i \sim N(f_0^i, \sigma_i^2)$

W is measurement cov. matrix.
Most common case: heteroskedastic uncorrelated measurement error:

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) \equiv W_{ij} = \delta_{ij} \sigma_i^2$$

Measurement error model:

$$\text{data} = \text{latent value} + \text{meas. error}$$
$$\underline{y_0} = \underline{f_0} + \underline{\varepsilon}$$

$$\underline{\varepsilon} \sim N(\underline{0}, \underline{\underline{W}})$$

(mean-zero Gaussian noise)

(cov matrix W)

[Derivation as the sum of two GPs at the observed times: Probabilistic Generative Model]

GP of intrinsic curve

$$f(t) \sim GP(m(t) = c,\ k(t, t'))$$

f_o = latent fn. at observed times t_o

GP Prior: $$f_o \sim N(\mathbb{1}c,\ K(t_o, t_o))$$

GP of measurement error:

y_o = measurements (with error) of f_o at times t_o

$$y_o | f_o \sim N(f_o, W)$$

Same as:

$$y_o = f_o + \varepsilon$$

(mean-zero error)
$\varepsilon \sim N(0, W)$

GOAL: Need Joint Dist. of data and latent fn.

$$\begin{pmatrix} y_o \\ f_* \end{pmatrix} \sim N\left(\begin{bmatrix} ? \\ ? \end{bmatrix}, \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix} \right)$$

(y_o: data we've seen; f_*: want to infer true curve for unobserved times)

Then can calc fn. (posterior) prediction at unobserved points

$$f_* | y_o \sim N(\mathbb{E}[f_* | y_o],\ Var[f_* | y_o])$$

Using conditional properties of MV Gaussian as before.

$$\begin{bmatrix} y_o \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbb{E}[y_o] \\ \mathbb{E}[f_*] \end{bmatrix}, \begin{bmatrix} Cov[y_o, y_o] & Cov(y_o, f_*) \\ Cov[f_*, y_o] & Cov[f_*, f_*] \end{bmatrix} \right)$$

< just fill in the gaps by calculating / substitute $y_o = f_o + \varepsilon$

Intrinsic / Latent Process $f_0 \sim N(1c, K(t_0, t_0))$

Measurement Process $y_0 | f_0 \sim N(f_0, W)$

$y_0 = f_0 + \varepsilon$ $\varepsilon \sim N(0, W)$ (meas. error)

To find cov. submatrices, apply bilinearity of covariance

$Cov(y_0, y_0) = Cov(f_0, f_0) + Cov(\varepsilon, \varepsilon) + 2Cov(f_0, \varepsilon)$

$Cov(f_0, f_0) = K(t_0, t_0)$ (GP of intrinsic curve)

$Cov(\varepsilon, \varepsilon) = W$ (measurement noise)

(the two processes are indep. → uncorrelated)

$2Cov[f_0, \varepsilon] = 0$

$\therefore \quad Cov[y_0, y_0] = K[t_0, t_0) + W$ //

Use joint of latent f at observed and unobserved times

$$\begin{pmatrix} f_0 \\ f_* \end{pmatrix} \sim N\left(\begin{bmatrix} 1c \\ 1c \end{bmatrix}, \begin{bmatrix} K(t_0,t_0) & K(t_0,t_*) \\ K(t_*,t_0) & K(t_*,t_*) \end{bmatrix} \right)$$

To derive similar arguments for

$Cov[y_0, f_*] = Cov[f_0, f_*] + Cov[\varepsilon, f_*] = K(t_0, t_*) + \underline{0}$

$Cov[f_*, f_*] = K(t_*, t_*)$

remember: K is a matrix

Fill out cov. matrix:

$$\begin{pmatrix} y_0 \\ f_* \end{pmatrix} \sim N\left(\begin{bmatrix} 1c \\ 1c \end{bmatrix}, \begin{bmatrix} K(t_0,t_0) + W & K(t_0,t_*) \\ K(t_*,t_0) & K(t_*,t_*) \end{bmatrix} \right)$$

now can compute fn. prediction at unobserved points:

Posterior Predictive (conditional on data y_0)

$$f_* | y_0 \sim N(\mathbb{E}[f_* | y_0], \text{Var}[f_* | y_0])$$

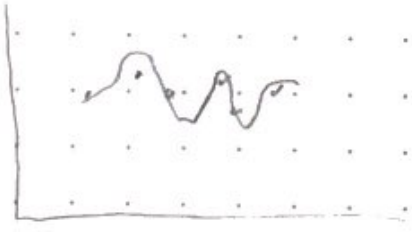
Using MV Gaussian conditional properties:

Posterior predictive mean

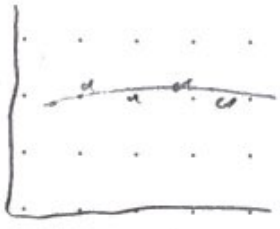
$$\mathbb{E}[f_* | y_0] = 1c + K(t_*, t_0)[K(t_0, t_0) + W]^{-1}(y_0 - 1c)$$

Posterior predictive variance/covariance

$$\text{Var}[f_* | y_0] = K(t_*, t_*) - K(t_*, t_0)[K(t_0, t_0) + W]^{-1} K(t, t_*)$$



ignore meas error: overfitting



include meas error → looks more like an actual separation

(see slides)

Now we know how to do ① PREDICTION,
How do we do ② MODEL SELECTION?

Tuning hyperparameters (A, τ)

Recall our probabilistic generative model:
Intrinsic/latent GP process
(make explicit dependence on hyperparameters)

$$f(t) \sim GP(m(t) = c,\ k_{A,\tau^2}(t,t'))$$

kernel: $k_{A,\tau^2}(t,t') = A^2 \exp(-|t-t'|^2/\tau^2)$

$$f_0 \sim N(1c,\ K_{A,\tau^2}(t_0,t_0))$$

Gaussian measurement process $y_0 = f_0 + \varepsilon$ $\quad \varepsilon \sim N(0,W)$

(if $x \sim N(\mu_x,\Sigma_x)$, $y \sim N(\mu_y,\Sigma_y)$ then $\begin{bmatrix}x\\y\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_x\\\mu_y\end{bmatrix}, \begin{bmatrix}\Sigma_x & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_y\end{bmatrix}\right)$)

Integrating out the latent fn. $f(t)$ (or use addition of MVN R.V.s) gives us the marginal likelihood:

$$P(y_0|A,\tau^2) = \int P(y_0|f_0) \times P(f_0|A,\tau^2)\, df_0$$

(cf. $\int P(y_0,f_0|A,\tau^2)\, df_0 = \int P(y_0|f_0;A,\tau^2) P(f_0|A,\tau^2)\, df_0$)

meas. error latent GP

(cf. $\int P(y_0,f_0)\, df_0$. joint normal $= \int N\left(\begin{pmatrix}y_0\\f_0\end{pmatrix}; \begin{pmatrix}1c\\1c\end{pmatrix}, \begin{pmatrix}C_{yy} & C_{yf}\\ C_{fy} & C_{ff}\end{pmatrix}\right) df_0$ marginalize $= N(y_0|1c, C_{yy} - C_{fy}C_{ff}^{-1}C_{yf})$)

$$= \int N[y_0|f_0,W] \times N[f_0|1c, K(t_0,t_0)]\, df_0$$

$$L(A,\tau^2) = P(y_0|A,\tau^2) = N[y_0|1c, K_{A,\tau^2}(t_0,t_0) + W]$$

- W = meas. error covariance
- K_{A,τ^2} = GP covariance

Which we can optimize (max likelihood) or specify a prior on (A,τ) and sample from posterior.

results for this example on slides
(!) find for this case, its not actually that sensitive to hyperparameters

remember
we had 4 images shifted by a time delay
each has indep. noise, how do we deal w/ this using GP?

Case study: Construct Bayesian Model

One latent fn. modelled as a draw from a GP

$g(t) \sim GP(c, k(t,t'))$ $k(t,t') = A^2 \exp[-(t-t')^2/\tau^2]$

Data of each image are measurements of $(\Delta t_i, \Delta m_i)$-shifted copies of this latent function

For image S1 w/ observations y_1 at times t_1:

$y_1(t_1) = g(t_1) + \varepsilon_1$ $\varepsilon_{1,j} \sim N(0, \sigma^2_{1,j})$

(j indexes the observations at times $t_{1,j}$)

Images S#i = 2,3,4 w/ observations y_i at times t_i w/ (unknown) time delays and magnitude shifts (relative to S1): $(\Delta t_i, \Delta m_i)$

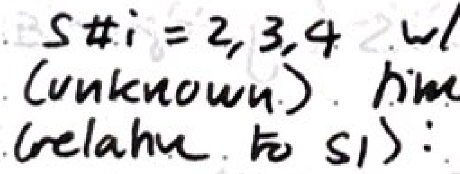
$i = 2,3,4$ $y_i(t_i) = \Delta m_i + g(t_i - \Delta t_i) + \varepsilon_i$

$\varepsilon_{i,j} \sim N(0, \sigma^2_{i,j})$

Sample Posterior: $P(\Delta m, \Delta t, A, \tau^2 \mid D)$

slides - results for sample w/ metropolis-within-gibbs
tune params for sampling as needed

marginals for time delays: Δt_2 Δt_3



gives best estimate for each time delay

shifting by estimated Δt, Δm, datasets line up pretty well as hoped!

Lecture 22

14.3.25

gelman BDA ch 21, 5

Today: Finishing Gaussian processes, starting hierarchical Bayes / Probabilistic Graphical Models

Gaussian Processes cont.

Previously: used "squared exponential kernel"

→ $\text{Cov}[f(t), f(t')] = k(t,t') = A^2 \exp(-|t-t'|^2/\tau^2)$

$f|A,\tau \sim N(\mu, K)$

(characteristic amplitude; characteristic timescale of fluctuations)

this kernel gives very smooth curves (infinitely differentiable)

also has properties: (diagonals tend to stationary variance)

Stationary: $k(t,t')$ time translation invariant to $t \to t+k$, $t' \to t'+k$ (i.e. only depends on separation t, t')

Symmetric: $k(t,t') = k(t',t)$

OTHER COV. FUNCTIONS:

"Ornstein Uhlenbeck Process" (damped random walk) (often used to model fin that follow)
Exponential covariance function

$$k(t,t') = A^2 \exp(-|t-t'|/\tau)$$

Long-run dist. of stochastic differential eq.

$$df(t) = \tau^{-1}[\mu - f(t)]dt + \sigma dW_t$$

(mean-reversion drag term; random walk)

$A = \tau\sigma^2/2$

(long-term mean; volatility; mean-reversion timescale)

paths

more jaggyed → is smooth choice squared exponential

→ is wiggly more jaggyer use this (depends on what you want for your iserae)

this kernel is everywhere continuous but not differentiable

"Matern class" of cov functions:

somewhere inbetween smooth & jaggyd

$\nu = 1/2$ special case of Matern kernel:

$$k_{\text{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{\ell} \right)$$

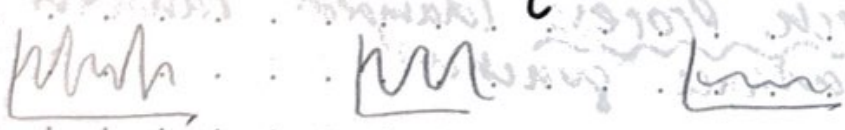
with +ve parameters ν, ℓ, where K_ν is a modified Bessel fn.

When $\nu = p + 1/2$ is half-integer; exponential × p-polynomial

$$k_{\nu=p+1/2}(r) = \exp\left(-\frac{\sqrt{2\nu} r}{\ell} \right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\frac{\sqrt{8\nu} r}{\ell} \right)^{p-i}$$

"Periodic cov. functions" (see ex sheet 3)

e.g. $k(t,t') = A^2 \exp\left(-\frac{2r^2}{\ell^2} \sin^2(\pi(t-t')/T) \right)$



end of GP

(board)

PROBABILISTIC GRAPHICAL MODELS

(conditional independence)

(Depicting complex probability distributions.)

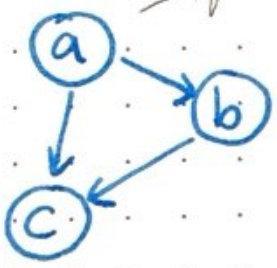
Suppose a, b, c are r.v.s with a joint pdf:

$$P(a,b,c) = P(c|a,b)P(a,b)$$
$$= P(c|a,b)P(b|a)P(a)$$

(general factorisation)

Directed Acyclic Graph:

(parent)



nodes are connected by one-way edges that do not form any cycles

DAG

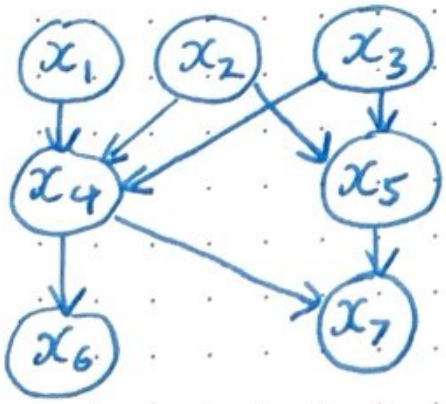
why care about dist'g rvs: everything in bayesian statistics is a dist. of rvs cf care about structured joint, posterior, etc.

Generally, if $\vec{x}$ is a collection of k r.v.s, then given a DAG, the factorisation is written as

if x_k has no parents then just $P(x_k)$

$$P(\vec{x}) = \prod_{k=1}^{k} P(x_k \mid \text{Parents of } x_k)$$

E.g.



$$P(x_1,\ldots,x_7) = P(x_1)P(x_2)P(x_3)$$
$$\times P(x_4 \mid x_1, x_2, x_3)$$
$$\times P(x_5 \mid x_2, x_3)P(x_6 \mid x_4)$$
$$\times P(x_7 \mid x_4, x_5)$$

we call this CONDITIONAL INDEPENDENCE

(the lack of connections implies some model structure!)

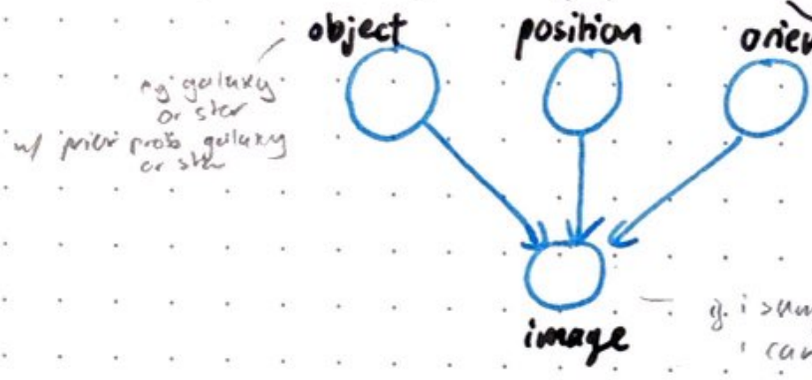
fact it isn't fully connected implies conditional independence (structure)

imagine you want to simulate an image with lots of galaxies in it

Generative Models (slides)

rv on sky

Causal process for generating images

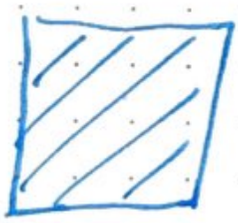


e.g. i sample joint dist of (object, pos, orientation) i can produce an image of galaxies/stars

can think of this joint dist as a graph (:)

Notation (board)

open node = latent parameter / unobserved data



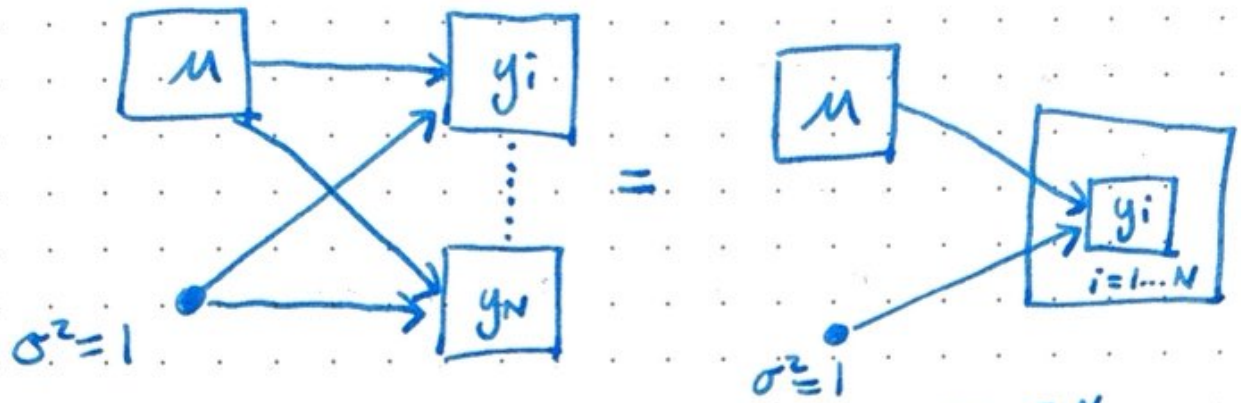
shaded node = observed parameter or data (conditioned on)

filled dot = fixed and known constant

plate: N independent replications of what's inside

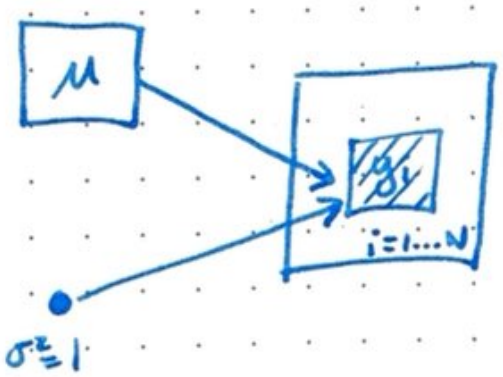
Eg. Gaussian i.i.d data $y_i \overset{iid}{\sim} N(\mu, \sigma^2 = 1)$

$i = 1, \ldots, N$



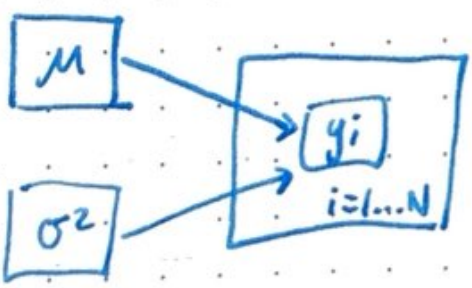
$$= P(\mu, \vec{y}) = P(\mu)\left[\prod_{i=1}^{N} P(y_i|\mu)\right]$$

If y observed, (shaded)



$$\propto P(\mu|\vec{y}) \propto P(\mu)\left[\prod_{i=1}^{N} P(y_i|\mu)\right]$$

Eg. Both μ, σ^2 unknown



$$= P(\mu, \sigma^2, \vec{y})$$

$$= \underbrace{P(\mu)P(\sigma^2)}_{P(\mu, \sigma^2)}\left[\prod_{i=1}^{N} N(y_i|\mu, \sigma^2)\right]$$

Conditional Independence and Probabilistic Graphical Models

Let a, b, c be r.v.s

(Marginal) Independence

$$P(a,b) = P(a)P(b) \rightarrow a \perp b$$
$$\rightarrow a \perp b \mid \emptyset$$

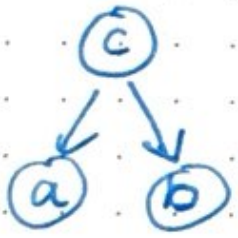
(not conditioning on anything)

Conditional Independence

$$P(a,b|c) = P(a|c)P(b|c) \rightarrow a \perp b \mid c$$

i.e. $P(a|b,c) = P(a|c)$ ← if I know c then b gives me no more info about a

Eg.

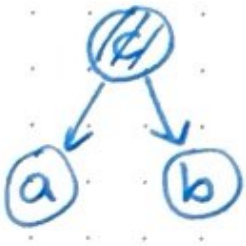


$$P(a,b,c) = P(c)P(a|c)P(b|c)$$

$$P(a,b) = \int P(c)P(a|c)P(b|c)\,dc \neq P(a)P(b)$$

$$\rightarrow a \not\perp b \mid \emptyset$$

"a is not indep of b marginally"

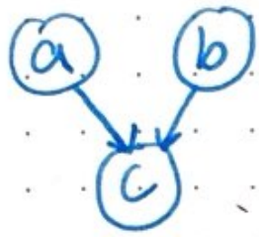


$$P(a,b|c) = \frac{P(a,b,c)}{P(c)} = P(a|c)P(b|c)$$

$$\rightarrow a \perp b \mid c$$

a is indep of b conditional on c

Eg.

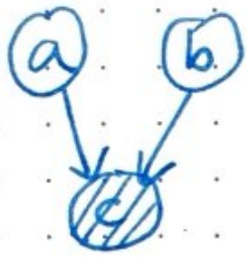


$P(a,b,c) = P(a)P(b)P(c|a,b)$

$P(a,b) = P(a)P(b)\int P(c|a,b)\,dc$

$= P(a)P(b) \rightarrow a \perp b \mid \emptyset$

"marginal independence between a & b"



$P(a,b|c) = \frac{P(a,b,c)}{P(c)} \neq P(a|c)P(b|c)$

$\rightarrow a \not\perp b \mid c$

i.e. by observing c, the child of a & b, our inferences a, b are now correlated

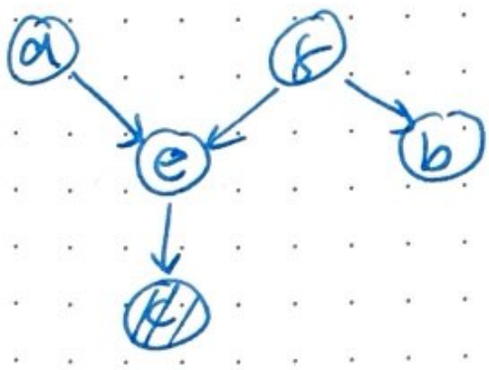
"a is no longer indep. of b if you condition on c"

D-separation Example

(slides)

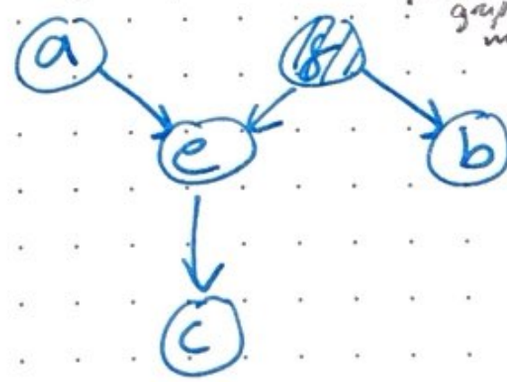
want to test whether nodes are indep given other nodes

reading independence properties from probabilistic graphical models



$a \not\perp b \mid c$

path a->b is not blocked so $a \not\perp b \mid c$



$a \perp b \mid f$

path a->b blocked so $a \perp b \mid f$

D-separation (slides)

* A, B and C are non-intersecting subsets of nodes in a directed graph

* A path from A to B is blocked if it contains a node such that either

a) The arrows on the path meet either → → head-to-tail or tail-to-tail at the node, ← → and the node is in the set C, or

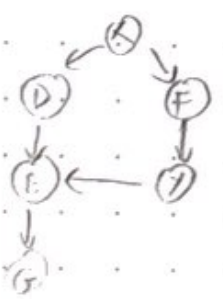
b) The arrows meet head-to-head at the → ← node, and neither the node, nor any of its descendants are in the set C.

* If all paths from A to B are blocked, A is said to be d-seperated from B by C.

* If A is d-seperated from B by C, the joint over all variables in the graph satisfies $A \perp B \mid C$

example → see slides
(very brief, didn't dwell on this in lecture)

e.g.



$X \not\perp Y$

$X \perp Y \mid F$

$X \not\perp Y \mid F, G$

way to think about statistical model: data generating process: how do I generate samples from parameters? → gives sampling distribution → observe data → likelihood function → + prior → bayesian inference

Lecture 23

17.3.25

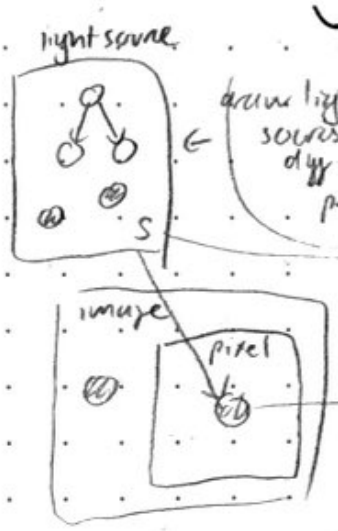
first step, this is what our graphs can describe → story starts w DGP

gelman BDA

(why we like models / why they are so useful / fundamental to our inference process.)

Today: Hierarchical Bayes / Bayesian Model selection

eg generative model (slides)



draw light sources & diff types / properties to each source

(plate ⇒ S sources)

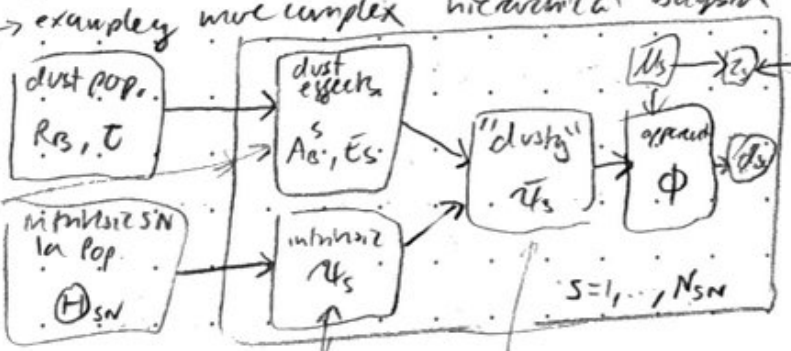
data generative model for astronomical images

each light source gives brightness to each pixel

PGM linked to hierarchical bayesian models

(all of mosey's research is basically hierarchical bayes)

Supernovae → example of more complex hierarchical bayesian model



(dist describing dust in galaxy)

correlated params that affect SN brightness → more realistically describe SN

(model intrinsic properties coming from population as if there were no dust)

combined effect intrinsic SN & dust

dust makes SN look dimmer & redder

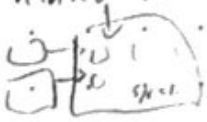
model spectral energy distribution SED — SED = model for brightness (flux) as fn of time & wavelength

now diff SN can have diff time series as fn of wavelength - trace it to variations in this region as fn of time

slide 4 describes 1 SN, slide 5 describing N SN (plate)

, of params

things outside plate describes pop distribution / cosmological parameters (things not related to individual SN) and things inside plate describes things related to individual supernovae



(Jou's pets not important)

while drugs example: big complicated hierarchical bayesian model. — see slides if you are interested

Common Problems in Astronomy

(slides)

i.e. not just one supernova, one galaxy etc → want to learn about population

* Want to learn about a population of objects from a finite sample of individuals, each measured with error

* Observed data is actually a combination of uncertain astrophysical & instrumental & selection effects. Need to model them to infer the "intrinsic" properties of the object or population of objects ("deconvolve")

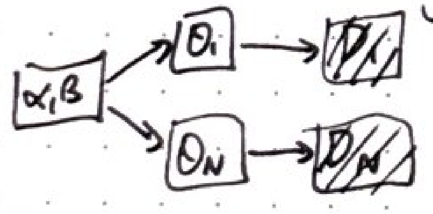
What is Hierarchical Bayes?

(slides)

"simple" Bayes: $D|\theta \sim \text{Model}(\theta)$

Posterior (Bayes' theorem): $P(\theta|D) \propto P(D|\theta)P(\theta)$

Hierarchical Bayes:



θ_i parameter of individual

α, β hyperparameters of population

$D_i|\theta_i \sim \text{Model}(\theta_i)$

$\theta_i|\alpha, \beta \sim \text{PopModel}(\alpha, \beta)$

Joint posterior:

$$P(\{\theta_i\}, \alpha, \beta | \{D_i\}) \propto \left[\prod_{i=1}^{N} P(D_i|\theta_i) P(\theta_i|\alpha, \beta) \right] P(\alpha, \beta)$$

build up complexity by layering conditional probabilities

diff hierarchies of params — e.g. population level parameters (pop. dist. of host galaxy dist)

object level parameters e.g. property of individual SN

Forward model:

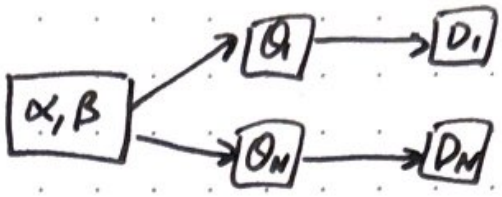
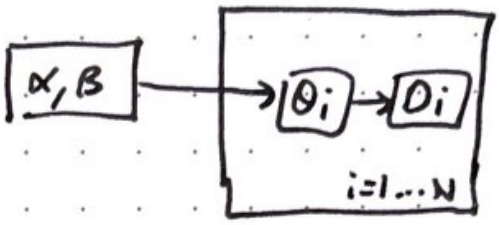


Plate notation:
(loop over individuals in sample)



Advantages of Hierarchical Bayesian Models

- Common problem in astronomy: infer properties of population from finite sample of individuals with noisy measurements
- Incorporate multiple sources of randomness & uncertainty as "latent variables" with distributions underlying the data
- Express structured probability models adapted to data-generating process ("forward model")
- Bayesian: Full (non-gaussian) probability distribution = global, coherent quantification of uncertainties
- Completely explore & marginalize posterior trade-offs / degeneracies between parameters / hyperparameters

(slides)

Simplest Hierarchical Bayesian / Multi-level Model: "Normal-Normal" for standard candle mag.

$s = 1, \ldots, N$

Level 1: Pop. distribution of latent variables (absolute mags)

"population dist / prior"

$$M_s \sim N(M_0, \tau^2)$$

M_s: latent variables; M_0, τ^2: hyperparams (pop. means & variance)

Level 2: Measurement error process

"measurement likelihood"

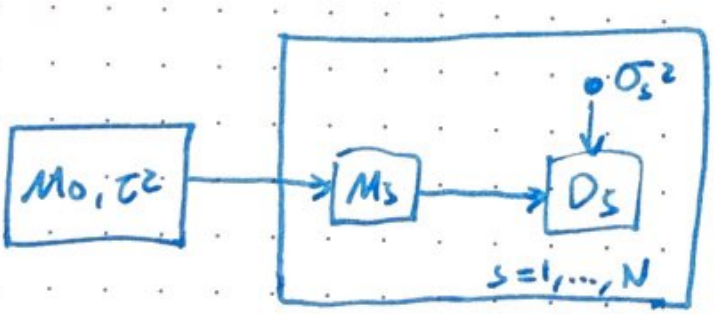
$$D_s | M_s \sim N(M_s, \sigma_s^2)$$

D_s: measurements (data); σ_s^2: heteroskedastic meas. error variance (known)

Joint probability density of data, latent variables, hyperparameters

"basically PGM for this joint"

$$P(\{D_s\}, \{M_s\}, H = \{M_0, \tau^2\})$$



joint factors into conditional and marginal pdfs based on model assumptions

Joint probability density of <u>all</u> the things: data, latent variables, hyperparameters

$$P(\{D_s\}, \{M_s\}, H) = \left[\prod_{s=1}^{N} P(D_s | M_s) P(M_s | M_0, \tau^2) \right] \times P(H)$$

measurement likelihood; population distribution/prior; hyperprior

Putting the Bayesian in Hierarchical Bayesian

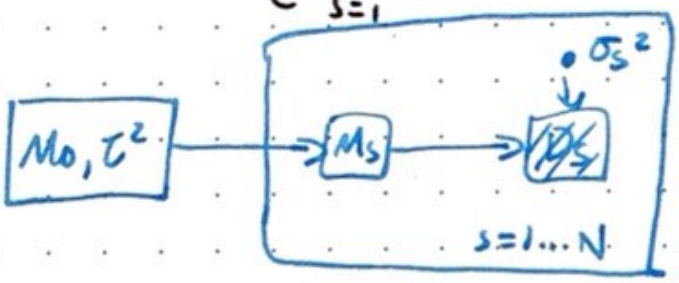
Joint posterior of all unknowns given the data

$$P(\{M_s\}, H \mid \{D_s\}) = \frac{P(\{M_s\}, H, \{D_s\})}{P(\{D_s\})}$$

← ignorable normalising const.
"ignore until next time"

(posterior on $N+2$ dim. parameter space)

$$P(\{M_s\}, H \mid \{D_s\}) \propto \left[\prod_{s=1}^{N} P(D_s \mid M_s) P(M_s \mid M_0, \tau^2) \right] \times P(H)$$



Hierarchical vs Regular Bayes

* Could regard as just a general Bayesian inference problem in a very high dim. parameter space, e.g.

$$\theta = \{M_1, \ldots, M_N, M_0, \tau^2\} = \{\underline{M}, M_0, \tau^2\}$$

$$P(\underline{\theta} \mid \underline{D}) \propto P(\underline{D} \mid \underline{\theta}) P(\underline{\theta})$$

$$P(\underline{\theta} \mid \underline{D}) \propto P(\underline{D} \mid \underline{M}) P(\underline{M} \mid M_0, \tau^2) P(M_0, \tau^2)$$

$$P(\underline{\theta} \mid \underline{D}) \propto \left[\prod_{s=1}^{N} P(D_s \mid M_s) P(M_s \mid M_0, \tau^2) \right] P(M_0, \tau^2)$$

helpful in bayesian computation if you can take advantage of the structure of the problem

* However, special hierarchical structure is useful for modelling, estimation & computation
* For large N, wouldn't want to do $N+2$ dim. Metropolis MCMC!

Gibbs sampling & Hierarchical Bayes

can develop algo for gibbs sampling that uses that structure in an efficient way

Utilises conditional independence structure of PGM/ posterior to derive conditional posterior densities

$$P(\{M_s\}, H \mid \{D_s\}) \propto \left[\prod_{s=1}^{N} P(D_s \mid M_s) P(M_s \mid M_0, \tau^2) \right] \times P(H)$$

Gibbs: use ① $P(\{M_s\} \mid H, \{D_s\})$ & ② $P(H \mid \{M_s\}, \{D_s\})$ instead to sample (need to plug in initial guesses first)

① For $s = 1, \dots, N$: Sample latent variables conditional on data and hyperparameters

$$P(M_s \mid H, D_s) \propto P(D_s \mid M_s) \times P(M_s \mid M_0, \tau^2)$$

SN indep!

(conditional independence) (gaussian) (gaussian)

product of 2 gaussians is gaussian → easy to sample from if we find mean, variance

② Sample hyperparameters from conditional on data and latent variables

gaussian

$$P(M_0, \tau^2 \mid \{M_s\}, \{D_s\}) = P(M_0, \tau^2 \mid \{M_s\})$$

(conditional independence)
↳ H conditionally indep of $\{D_s\}$ given $\{M_s\}$ → can read from graph!

$$= P(M_0 \mid \tau^2, \{M_s\}) P(\tau^2 \mid \{M_s\})$$

(gaussian) (inv-χ^2)

Reduces to the familiar posterior for unknown mean and variance of gaussian data (sheet 2 prob. 1 BDA 3.2-3.3)

- fix data (condition on D always)
- input some guesses for M_0, τ^2 and condition on that, sample latent variables
- turns out that if we condition on H, M of SN 1 is indep to M of SN 2, 3 etc.

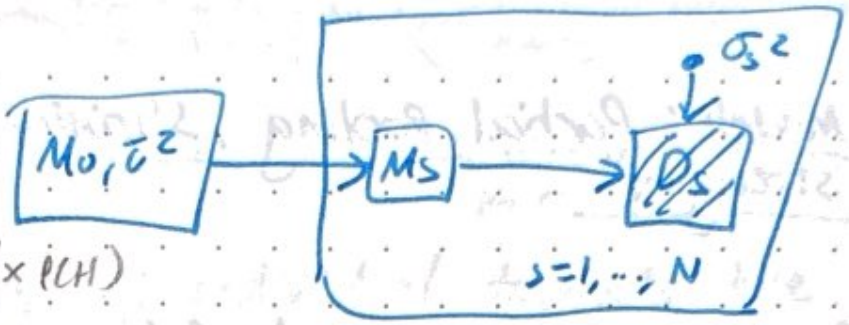
① so simplifies to two terms rather than N+2 terms which depend on s so can loop through each s which is easy to sample from b/c gaussian

② turns out if we read conditioning relationships of graph we don't have to worry about data, only these N factors but those look like gaussian which itself factors into a gaussian & inv-χ^2 → much easier to sample from

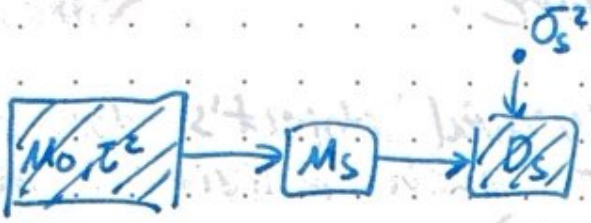
(w unknown M_0, σ^2?)

Full posterior:

$P(\{M_s\}, H | \{D_s\}) \propto \left[\prod_{s=1}^{N} P(D_s|M_s) P(M_s|M_0,\tau^2)\right] \times P(H)$



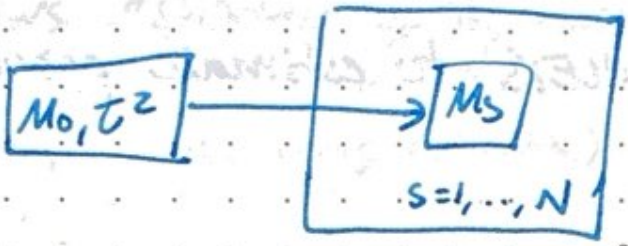
(1)



$P(M_s|M_0,\tau^2,\{D_s\}) \propto P(D_s|M_s) \times P(M_s|M_0,\tau^2)$

$\propto N(D_s|M_s,\sigma_s^2) \times N(M_s|M_0,\tau^2)$

(2)



← hyperprior

$P(M_0,\tau^2) = P(M_0)P(\tau^2) \propto 1$

$P(M_0) \propto 1$, $P(\tau^2) \propto 1$

$P(M_0,\tau^2|\{M_s\},\{D_s\}) = P(M_0,\tau^2|\{M_s\})$

$= P(M_0|\tau^2,\{M_s\}) P(\tau^2|\{M_s\})$

gaussian · inv χ^2

(similar to sheet 2 problem 1)

↙ "think about this on your own"

Hyperprior $P(M_0,\tau^2) \propto 1$

Draw from $\tau^2|\{M_s\} \sim \text{Inv-}\chi^2\left(N-3, \frac{(N-1)}{(N-3)} s^2\right)$

$M_0|\tau^2;\{M_s\} \sim N(\bar{M}, \tau^2/N)$

$\bar{M} \equiv \frac{1}{N}\sum_{s=1}^{N} M_s \qquad s^2 \equiv \frac{1}{N-1}\sum_{s=1}^{N}(M_s - \bar{M})^2$

---

e.g. 100 SN w/ abs mag & distance measurements

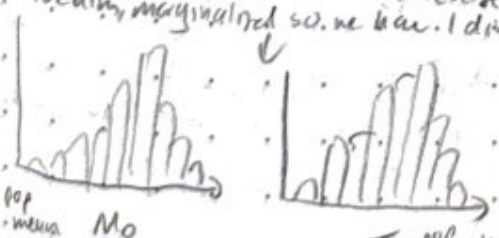
(can make histogram of obs data D_s)

$D_s = \hat{m}_s - \hat{\mu}_s$ — see slides

data with some error

gibbs sampling: sample $M_s|M_0,\tau^2$ then do the reverse: sample $M_0,\tau^2|M_s$

(is 102 dim, marginalized so we have 1 dim)



marginal posterior estimates

scatterplot

- posterior estimates closer to dashed line
- by doing hierarchical stages, the posterior estimates cluster to truth more precise (since original data contains measurement error w/ larger spread than truth)

individual estimates have narrower spread than posterior estimates (have learnt to deconvolve out measurement noise)

→ "partial pooling" phenomenon

HB models, improvement partial pooling + strength

(3 ideas)

HB Models: Partial Pooling, Shrinkage & "Borrowing of strength"

each sn: observe Di
find individual $p(D_i|\theta)$
find θ once for each

* Common sense procedure:
  - Analyse each individual object's data D_i separately and get each individual MLE: estimate (with error)
  - Plug in all {MLEi's} to estimate population hyperparameters

* **PROBLEM**: Each individual θ_i estimate may be unbiased but collectively give a biased estimate of population (e.g. variance) because of errors

* **SOLUTION**: Use HB to model & infer individuals & population simultaneously and get better estimates of both

notes to self: (attempt to develop intuition).

(HB full posterior) $\rightarrow P(M_s|M_0, \tau, \underline{D}) \propto P(D_s|M_s) P(M_s|M_0, \tau^2)$

individual estimate. ↑ pop dist acts as prior

say we want to infer M_s from and take peak as our best estimate of M_s:

* individual estimate just finds whatever M_s maximises prob of observing our data (MLE)
* but HB model full posterior incorporates prior on M_s so e.g. M_s may maximise $P(D_s|M_s)$ but if M_s is very unlikely this will sway our estimate. i.e. M_s through the posterior => balance between maximising prob of data but also probability of intrinsic abs mag M_s existing in the population in the first place

How is this diff to MAP?
=> point is HB gives $P(\tau|\{D_s\})$. so we can infer these hyper params instead of needing to chose them ourselves (I think)

(slides)

"often not unbiased but their biases helps you in a way"

Shrinkage Estimators

closer to truth on avg than considering individual estimate for each individual in isolation

* Bias estimator of individual towards the population
* Leads to overall lower MSE than individual unbiased estimators
* Allows "sharing of information" between individuals to improve overall estimation

example: sticks (brief explanation)

Shrinkage with Hierarchical Model

Return to our previous example:

Level 1: pop. dist. of latent variables — $M_s \sim N(M_0, \tau^2)$ — "pop. dist." / "prior"

Level 2: measurement error process — $D_s | M_s \sim N(M_s, \sigma_s^2)$ — "measurement likelihood"

(causing yearly differentials & our likelihood $P(D_s|M_s)$)

Individual MLE of s^{th} alone: $\hat{M}_s = D_s$ (unbiased)

Population dist. acts as prior

individual estimate comes from here

thinking pop dist as acting as prior

$$P(M_s | M_0, \tau^2, \{D_s\}) \propto P(D_s|M_s) P(M_s|M_0, \tau^2)$$

(pop dist.)

and pulls posterior estimate of individual closer to population mean estimate

τ controls how much we shrink our posterior estimate towards pop. mean estimate

pull/shrinkage controlled by population variance τ^2

theorem: from full posterior:

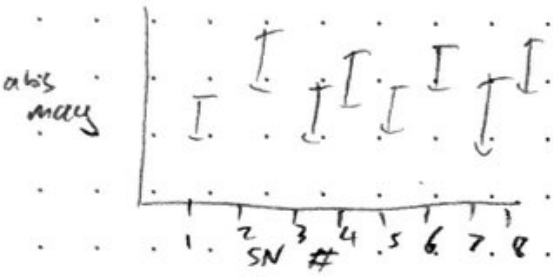
$$P(M_s, M_0, \tau^2 | \{D_s\}) \propto P(H) \left[\prod_{s=1}^{N} P(D_s|M_s) P(M_s|M_0, \tau^2) \right]$$

just ignore terms not in M_s like in simple gaussian example

$$P(M_s | M_0, \tau^2, \{D_s\}) \propto P(D_s|M_s) P(M_s|M_0, \tau^2)$$

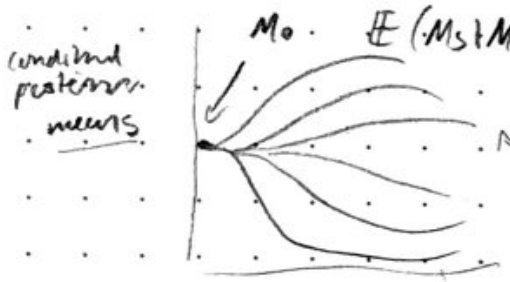
How does HB implement shrinking?

example (dataset of 8 SN)



$\hat{M}_s = D_s$

If you know M_0, τ of individual M_s turns out it depends on τ.



(one track for each SN, posterior mean as a function of τ)

If we can take τ to zero, converge to M_0

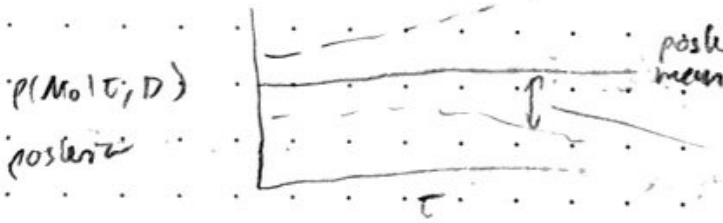
τ controls degree of shrinkage from M_0

(increase τ; wider range of M_s probable so more spread)

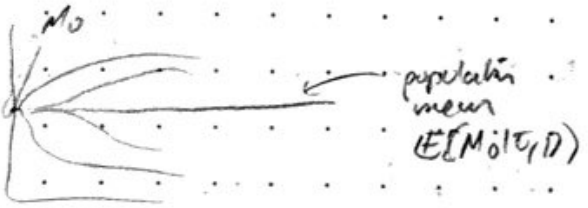
but we don't know M_0 -> how to estimate that from marginalized posterior

$$P(M_0, \tau | \underline{D}) = P(M_0 | \tau, \underline{D}) P(\tau | \underline{D})$$

marginalise out individual M_s



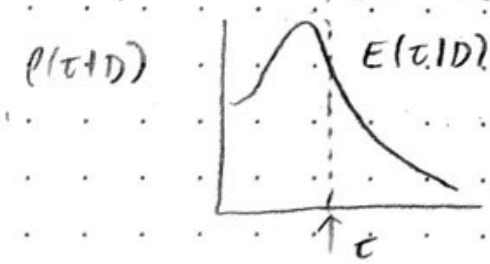
depends on τ



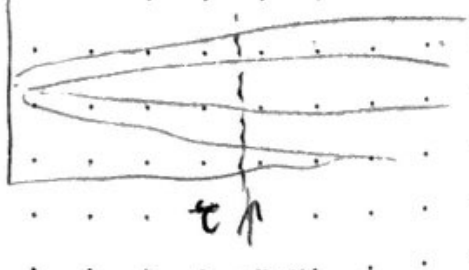
these curves shrink towards population mean with decreasing τ

hierarchical bayes gives the posterior density $P(\tau | D)$ -> way to estimate τ

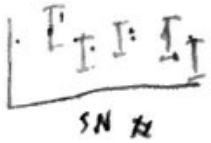
(don't necessarily have to take point estimate of τ, can do diff estimate?)



how can make at best trade-off between shrinking towards same pop. value vs allowing them to be independently estimated?
↳ this is at about $\tau = 0.134$



on avg HB estimate closer to truth than individual MLE



Lecture 24

recommended reading (will's book!)
(-> Mackay "Information theory, inference & learning algorithms" ch. 28

19.3.25

Today: Bayesian Model selection

Model Comparison is selection

Mackay: "2nd level of inference" 1st level: assume model true & fit that model to the data (MAP/ML) 2nd level is to compare models

- → No. of spectral lines in a (noisy) spectrum?
- → Clustering / mixture models - how many clumps?
- → Time series - curve fitting
  - Is there a trend?
  - Complexity / order / degree of best model (x^7 x^3 ?)
  - which GP kernel best explains the data?
- → Cosmology - standard (8-parameter) cosmological model vs more exotic (more parameters) models?
  - 6 in standard model (ΛCDM)
  but people (cosmologists) keep finding more exotic variants of the model (to try and figure out dark energy)
  but then we need to introduce more parameters
  ↳ which extra parameters are warranted by data (adding more params will always give better fit but how do we know if adding extra params is just fitting to noise or actually warranted by data?) -> hope to answer this w/ bayesian model comparison

Bayesian Model Comparison

Parameter estimation: posterior on parameters

$$P(\theta_1|D,M_1) = \frac{P(D|\theta_1,M_1)P(\theta_1|M_1)}{P(D|M_1)}$$

Model selection: posterior on models

$$P(D|M_1) = \int P(D|\theta_1,M_1)P(\theta_1|M_1)\,d\theta_1$$

(evidence or marginal likelihood)

needs proper prior!

Posterior odds ratio: between two competing models

$$\frac{P(M_1|D)}{P(M_2|D)} = \underbrace{\frac{P(D|M_1)}{P(D|M_2)}}_{BF} \times \frac{P(M_1)}{P(M_2)}$$

↖ Prior odds

prior-preference for model 1 vs model two (scientifically based) or if ignorant equal probabilities

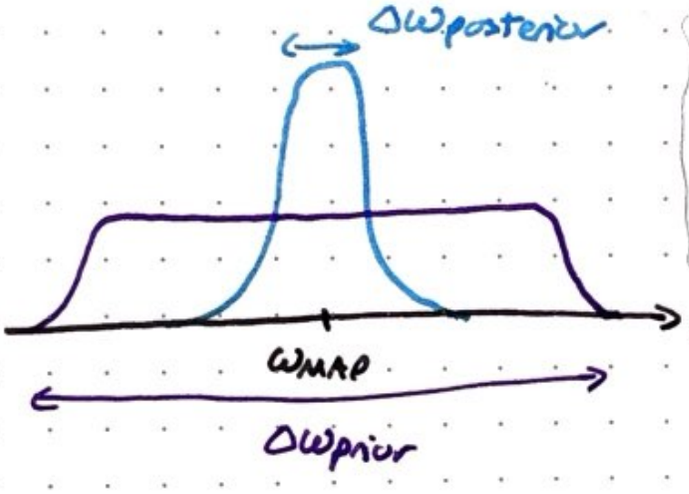
$$\text{Bayes Factor}_{12} = \frac{P(D|M_1)}{P(D|M_2)}$$

(let's develop some intuition)

For a given model with a single parameter, w, consider the approximation

$$p(D) = \int P(D|w)p(w)\,dw \simeq P(D|w_{MAP}) \frac{\Delta w_{posterior}}{\Delta w_{prior}}$$

where the posterior is assumed to be sharply peaked.



cartoon approximation

approx. log. evid at w_{MAP}:

- $P(D) \simeq P(D|w_{MAP}) \cdot \frac{P(w_{MAP})}{P(w_{MAP}|D)}$

assume post, prior are gaussian

- $P(w) \simeq N(w_{MAP}, \Sigma_{prior})$
- $P(w|D) \simeq N(w_{MAP}, \Sigma_{post})$

plug in

- $P(D) \simeq P(D|w_{MAP}) \frac{\Sigma_{prior}^{-1/2}}{\Sigma_{post}^{-1/2}}$
- $\simeq P(D|w_{MAP}) \frac{\Delta w_{post}}{\Delta w_{prior}}$

Taking logarithms, we obtain

$$\ln P(D) \simeq \ln P(D|w_{MAP}) + \ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$$

negative

since we hope prior width always larger than posterior width

generalize:
With M parameters, all assumed to have the same ratio $\Delta w_{posterior}/\Delta w_{prior}$, we get

$$\ln P(D) \simeq \ln P(D|w_{MAP}) + M \ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$$

(term related to fitting data with best fit parameters)

term that is negative, proportional to no of parameters & ratio widths

negative

(OCCAM FACTOR)

models. adding more parameters will always increase likelihood of data under that model, but this increase in likelihood is not always warranted because we could just be overfitting to noise. Occam's factor (2nd term) penalizes overfitting

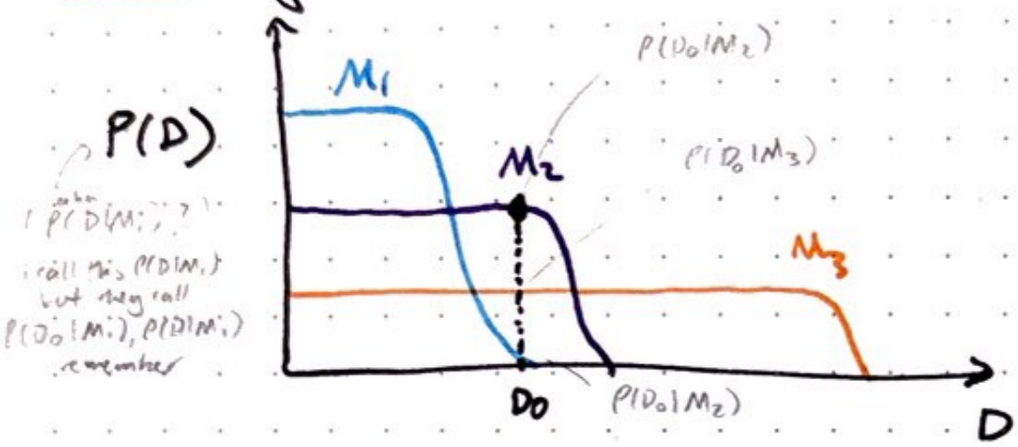
this term increases w parameters always

but this term penalizes you for adding more parameters

favors simpler models that explains the data

want to find sweet spot w enough parameters without overfitting

Matching data and model complexity



also P(D|Mi) but may call P(D0|Mi), P(D|Mi) remember

"more complex models can predict a greater variety of data sets"

- M1 can only predict limited amount of data (e.g. might be a linear model)
- M2 can predict slightly more complex data sets (e.g. might be a quadratic model)
- M3 e.g. if a neural network may be able to predict arbitrarily complex functions

but remember:

- P(D) is also a distribution over potential data sets
- the more complex a dataset a model can predict, the more spread out P(D) will be. It is normalised => giving particular dataset it will have a lower probability overall

=> so simpler models will have more limited range of data sets they can predict, but because of that, overall P(D) is higher

=> whereas complex models predict greater range but P(D) lower bc of that normalisation factor (parts any given one)

so this gives a good way to weigh fit of models to dataset vs their overall complexity

eg: for dataset D0. M1 is v. low probability. M1 describes D0 dataset. M3 slightly more probable but M2 has best balance between complexity of model (more complex than M1) but also not too complex. (M2)

recommend read ch 28 of Mackay

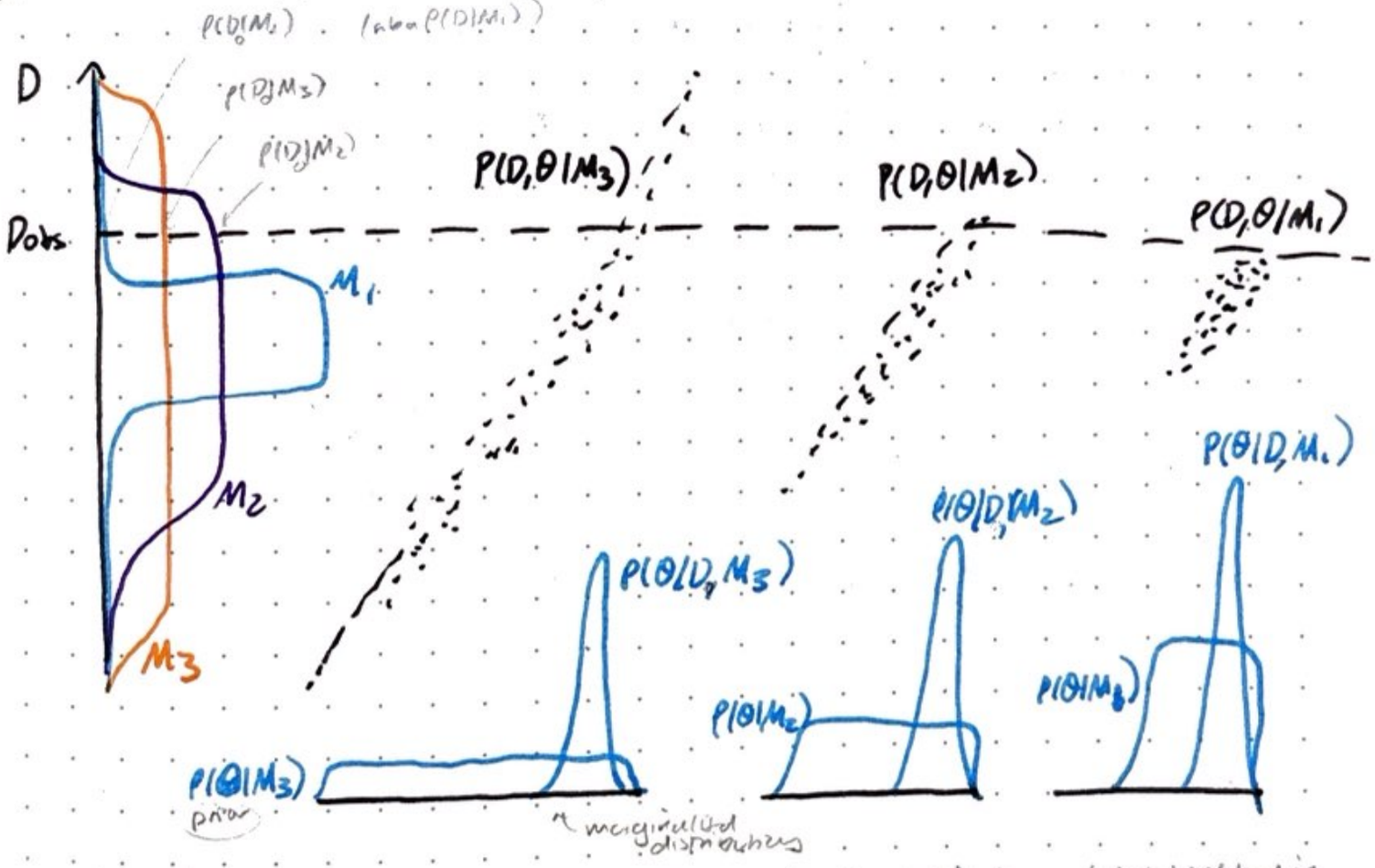
Marginalising Joint Distribution of Parameters and Data

hypothesis H

$(\omega \sim P(\omega|H)$ & $D \sim P(D|\omega,H) \rightarrow (D,\omega) \sim P(D,\omega|H))$

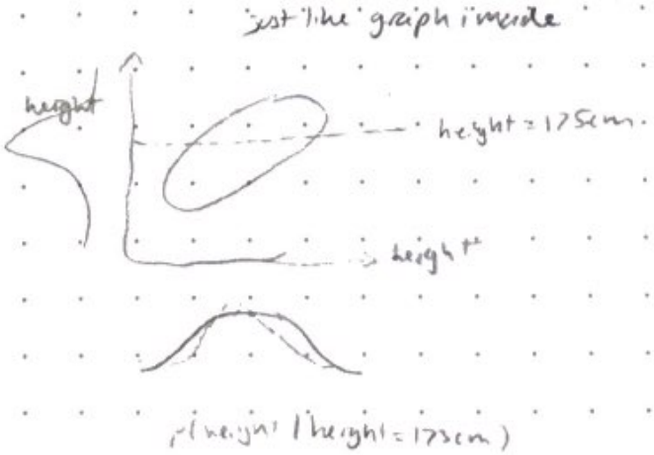
I'm going to call this model, M

$\theta \sim P(\theta|D)_M$ & $D \sim P(D|\theta,M) \rightarrow (D,\theta) \sim P(D,\theta|M)$



draw θ from prior then given θ draw D from sampling distribution (what I actually want is desired joint) → this gives θ and D pairs from the joint distribution e.g. $P(D,\theta|M_1)$ (data are not data points they are samples from joint "observed" data set there is just a single value shown by dotted line)

(draw from prior $\theta \sim P(\theta|M)$
given that draw from sampling dist $D \sim P(D|M,\theta)$
gives pairs θ, D from joint)



used to be directory IoA!

Interpreting the Bayes Factor / Evidence Ratio: The Jeffreys Scale

$$\Delta \ln E = \ln BF$$

these categories are up for debate / not rigorous

$\Delta \ln E < 1$	"not worth more than a bare mention"
$1 < \Delta \ln E < 2.5$	"significant"
$2.5 < \Delta \ln E < 5$	"strong to very strong"
$5 < \Delta \ln E$	"decisive"

How to calculate Evidence

$$P(D) = \int P(D|\theta) P(\theta) d\theta$$

- Analytic (really simple, nice problems) (as analytic integral)
- Laplace approximation
- Savage-Dickey Ratio (Nested Models) (see later)
- Monte-Carlo: probably slow if likelihood is peaked (likely slow: almost certainly strongly peaked if is at all informative over my data)

fundamentally because approx integral by sum

$$\theta_i \sim P(\theta_i) \qquad P(D) \approx \frac{1}{M} \sum_{i=1}^{M} P(D|\theta_i)$$

$i = 1, \ldots, M$

approximate with MC average likelihood

- Harmonic Mean Estimator (unstable) (has infinite variance)

$$\theta_i \sim P(\theta|D) \qquad P(D) \approx \left[\frac{1}{M} \sum_{i=1}^{M} P(D|\theta_i)^{-1} \right]^{-1}$$

MC arg is inverse then find inverse of that

- Nested Sampling

Calc. evidence via laplace approx:

(Recall: The Laplace Approximation)

Evidence $P(D) = \int P^*(\theta|D)\,d\theta$

Unnormalised posterior $P^*(\theta|D) = P(D|\theta)P(\theta)$

Find MAP estimate: $\theta_0 = \underset{\theta}{\operatorname{argmax}} \ln P^*(\theta|D)$

Taylor expansion:

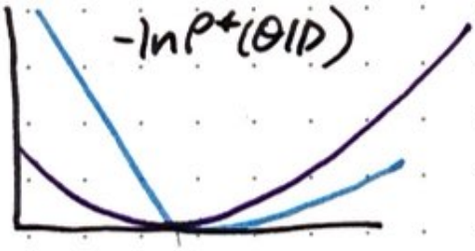
$$\ln P^*(\theta|D) \simeq \ln P^*(\theta_0|D) - \tfrac{1}{2}(\theta-\theta_0)^T A (\theta-\theta_0) + \dots$$

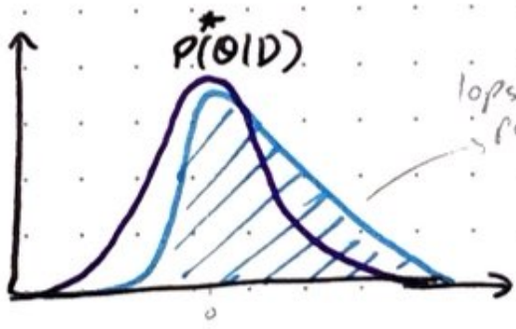
Hessian at mode: $A_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln P^*(\theta|D)\Big|_{\theta=\theta_0}$

$$P^*(\theta|D) \simeq P^*(\theta_0|D) \times \exp\left(-\tfrac{1}{2}(\theta-\theta_0)^T A (\theta-\theta_0)\right)$$

Not always accurate:

eg.





e.g. lopsided unnormalised posterior: laplace approx vs true approx

but if you trust your approximation you can do an integral over params & get this:

Evidence from Laplace Approximation:

$$P(D) = \int P^*(\theta|D)\,d\theta \simeq P^*(\theta_0|D) \times |2\pi A^{-1}|^{1/2} \int \mathcal{N}(\theta|\theta_0, A^{-1})\,d\theta$$

$P(D) \simeq P^*(\theta_0|D) \times \det(A/2\pi)^{-1/2}$

$P(D) \simeq P^*(\theta_0|D) \times \det(2\pi A^{-1})^{1/2}$

$$P(D) \simeq P(D|\theta_0) \times P(\theta_0) \times \det(A/2\pi)^{-1/2}$$

Hessian of negative log posterior

$A = -\nabla\nabla \ln P^*(\theta|D)$

unnormalised?

need to be able to differentiate posterior at least 2× to do this (so only work for tractable posteriors)

this implements occam's razor

Evidence implements Occam's Razor

someas / similar to analysis before

$$P(D) \approx P(D|\theta_0) \times P(\theta_0) \times \det(A/2\pi)^{-1/2}$$

Evidence — best fit likelihood — Occam factor

(?) or just det(A)?

$$\Sigma_{post} = A^{-1}$$ roughly

how can we see this?:

Suppose: $P(\theta) = N(\theta | \theta_{prior}, \Sigma_{prior})$ — suppose prior itself is gaussian

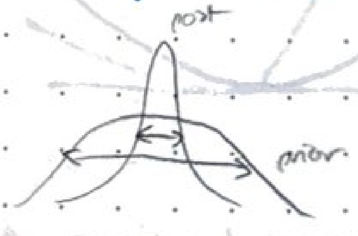
Occam Factor:

assume prior centred near MAP θ_0 so we can ignore exponential term? (chatgpt)

$$\propto |2\pi\Sigma_{prior}|^{-1/2} \times |2\pi\Sigma_{post}|^{1/2}$$

$$\propto \frac{|\Sigma_{post}|^{1/2}}{|\Sigma_{prior}|^{1/2}}$$

(ratio of posterior to prior width)



From inference book:

- factor by which our model's hypothesis space collapses when the data arrive
- magnitude of occam factor gives a measure of complexity of the model
- depends on both
  - → no of params in model
  - → prior prob model assigns to those params

more params → multiply factor by N (in wide prior/complex model) then smaller occam factor

notes to self:

$$\frac{P(D|M_1)}{P(D|M_2)}$$

← implements occam's razor for each, tradeoff between maximising data fit & minimising complexity of model so fairer/more accurate comparison?

Simple models concentrate probability around limited no of datasets

complex models much wider range of datasets but probability of a given dataset will be smaller i.e. even if both simple & complex model can predict data, complex will have much smaller prob of

Savage - Dickey Ratio

for nested models?
useful for if we have simple model nested within complex model

Suppose the parameters are ϕ, ψ and complex model M_1 reduces to simpler model M_0 when $\psi = 0$ (Nested)

e.g. cosmological params if we set new params in exotic model to zero we get 'simple' params (ΛCDM?)

And the prior is separable:

$$P(\phi, \psi | M_1) = P(\psi | M_1) P(\phi | M_1)$$

controls model complexity?

nuisance parameters (new params)

(models have free parameters in common that we're not directly interested in but the focus of our model comparison)

or e.g. linear as a special case of polynomial (constants = 0)

And the prior on ϕ is the same for each model:

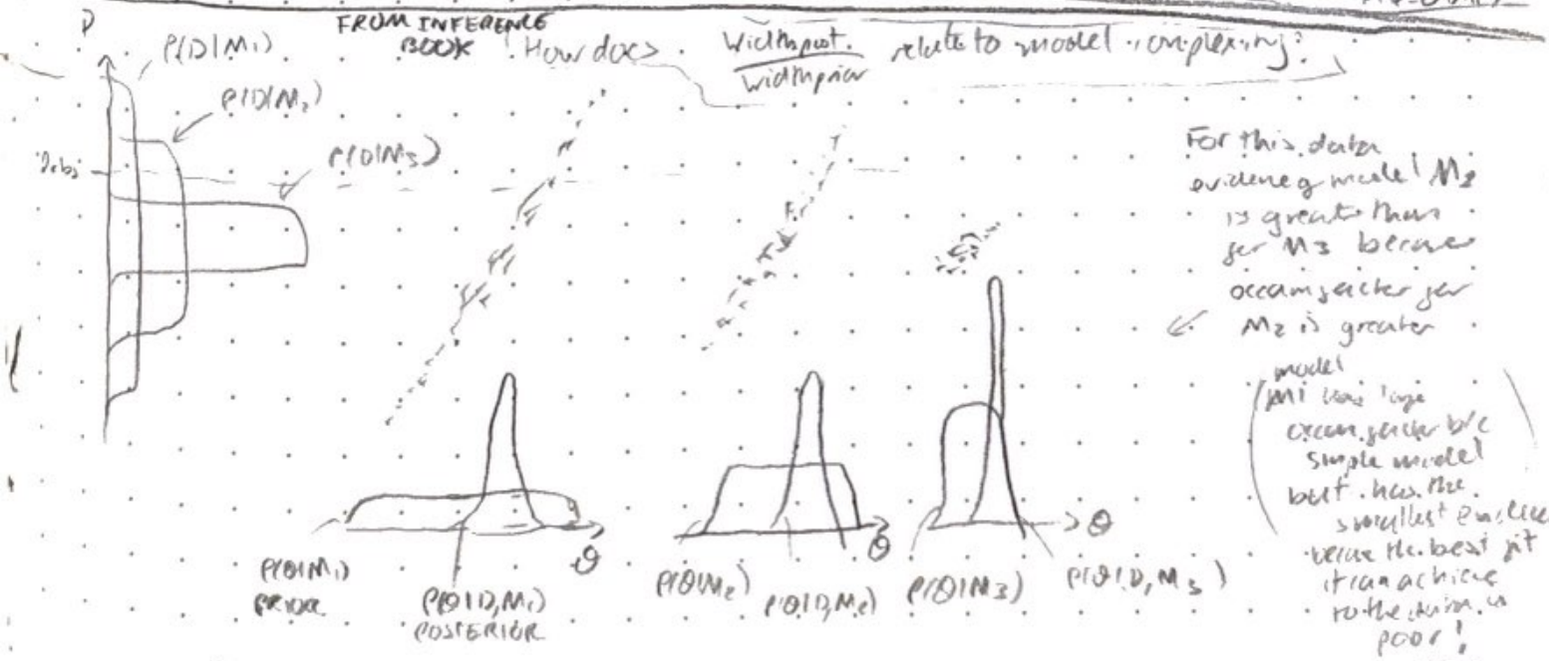
prior on nuisance parameters

$$P(\phi | M_1) = P(\phi | M_0)$$

Then the Bayes Factor reduces to:

$$B_{01} = \frac{P(D|M_0)}{P(D|M_1)} = \frac{P(\psi | D, M_1)}{P(\psi | M_1)}\bigg|_{\psi=0}$$

Show this:
- consider numerator $P(D|M_0)$
$= \int P(D|\phi, M_0) P(\phi|M_0) d\phi$
$= \int P(D|\phi, \psi=0, M_1) P(\phi|\psi=0, M_1) d\phi$
$= P(D|\psi=0, M_1)$
$= \frac{P(\psi=0|D, M_1) P(D|M_1)}{P(\psi=0|M_1)}$ (Bayes)
- now ratio becomes $B_{01} = \frac{P(\psi=0|D, M_1)}{P(\psi=0|M_1)}$



For this data evidence/model M_2 is greater than for M_3 because Occam factor for M_2 is greater

(model M_1 has large Occam factor b/c simple model but has the smallest evidence because the best fit it can achieve to the data is poor!)

Occam factor for M_3 small b/c wide prior for complex model

for evidence calculations but also for posterior sampling (useful bi-product!)

NESTED SAMPLING

- An algorithm designed to compute Bayesian evidence (John Skilling, 2004)
- Can be better than MCMC for multi-modal distributions
- Get weighted samples from posterior for free
- Evolve an ensemble of "live points" successively sampling prior volume above a likelihood-level constraint
- Perform evidence integral over contours of equal likelihood
- Uses statistical estimate of prior mass within likelihood-level

Want to calculate multi-dim. evidence integral

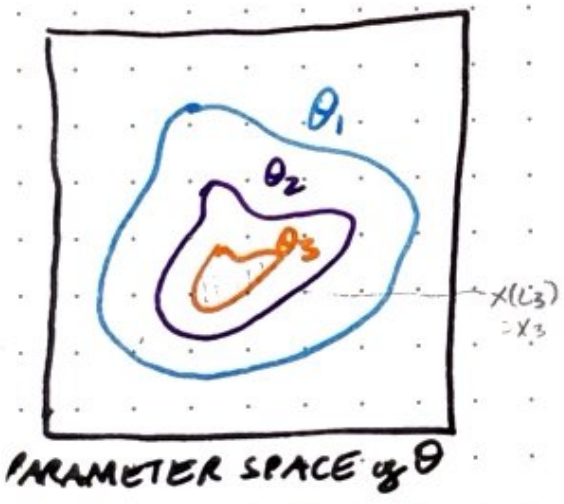
$$Z = \int L(\underline{\theta})\pi(\underline{\theta})\, d\underline{\theta}$$

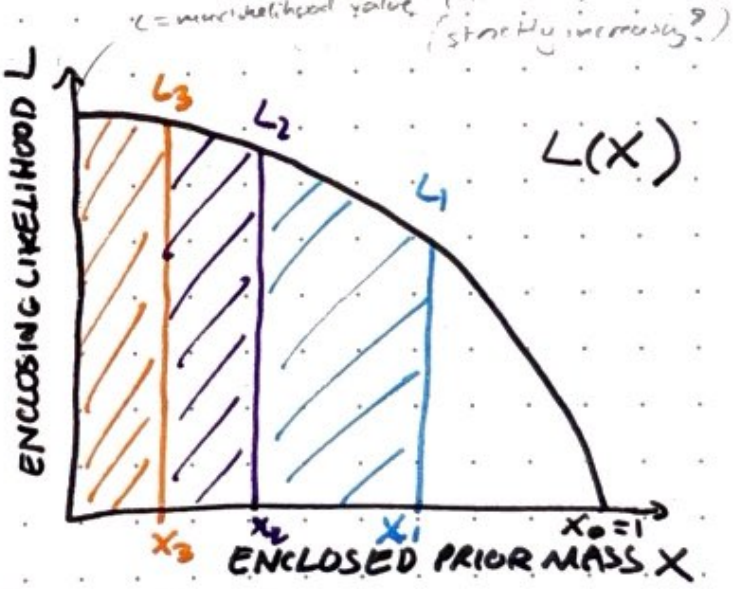
Define "prior mass enclosed above likelihood L^*"

$$X(L^*) = \int_{L(\underline{\theta})>L^*} \pi(\underline{\theta})\, d\underline{\theta}$$

Inverse: Likelihood at enclosed prior mass: $L(X)$

1D evidence integral: $Z = \int L(X)\, dX$





Nested likelihood contours sort to enclosed prior mass X.

e.g. 2D example take 3 random points from prior $\theta_1, \theta_2, \theta_3$ and calc. their likelihood, each will be associated with some likelihood contour each contour is associated with some percentage of the prior volume $X(L^*)$

SO can associate with each point a likelihood contour and also a prior mass enclosed above that likelihood

↳ if I can map from points in param. space to $L(X)$ I can turn a previously multi-dim. integral in param space to a 1-Dim integral in X from 0 to 1

how do we actually do this?

Nested sampling: How?

[NB: For the rest of this lecture assume prior is uniform box in parameter space (w Total integral 1)]

1. Begin by sampling N_{live} points $\{\theta_i\}$ from the prior ($X_0 = 1$), evaluate likelihood for each live point, initialise evidence $Z = 0$.

2. Find the live point with smallest likelihood L^*.

3. Estimate compression factor

$$t_i = X_i / X_{i-1} .$$

volumes X_i estimated statistically

$t_i \sim e^{-1/N_{live}}$

4. Accumulate evidence as a 1D integral

$$\Delta Z = L^* \times (X_{i-1} - X_i) = L^* \times (1 - t_i) X_i .$$

5. Sample a new live point from prior constrained to ~~$L > L^*$~~ $L > L^*$, compute its likelihood value.

(e.g. with MCMC or otherwise) reject ought not higher L

+ copy to dead points) simplest is to pick uniformly & repeat if $L < L^*$ (?)

6. Repeat steps 2–5 until convergence (i.e. evidence Z stops changing within some tolerance).

7. Add final evidence estimate of remaining live points

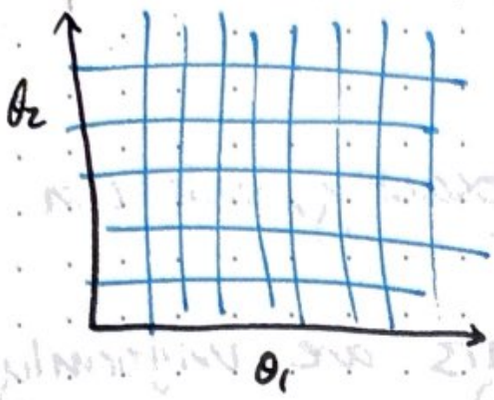
$$\Delta Z = \bar{L} \times X_{end} .$$

8. List of dead points give weighted posterior samples $\{\theta_i\}$ with

$$w_i \propto (1 - t_i) X_i$$

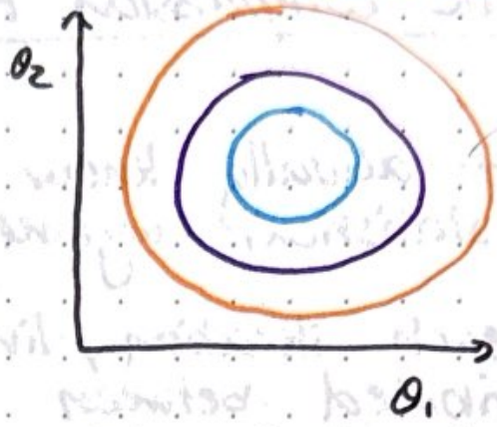
want to find Z: instead of directly into cubes & summing over L for each cube, 'combine' cubes w similar likelihood & perform easier 1D sum (only care about difference in evidence between contours?)

$Z \approx \sum L(\theta_1, \theta_2) \Delta\theta_1 \Delta\theta_2$

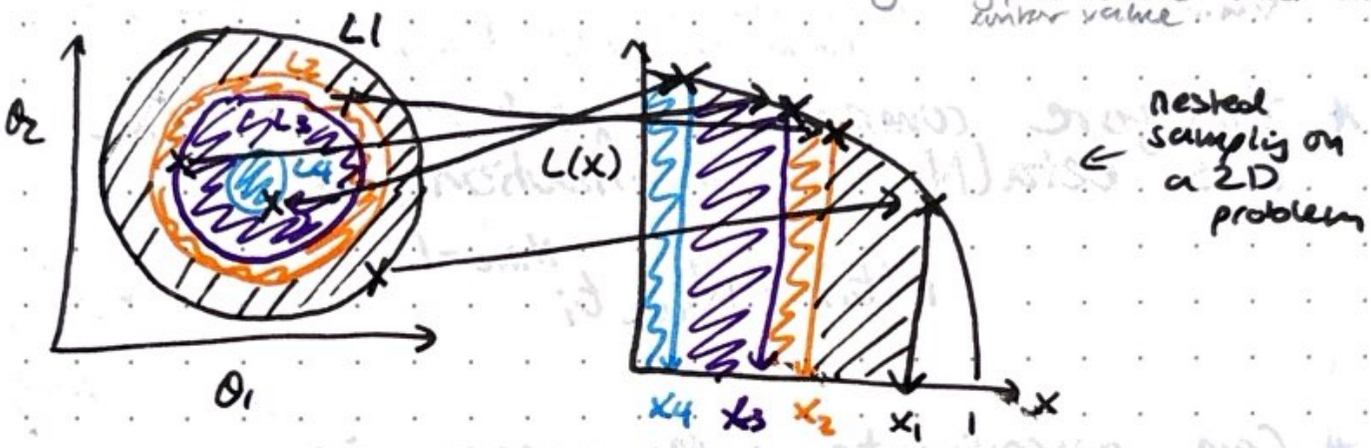


naive: sum over cubes and add them up
but in huge param spaces this is very computationally expensive

$Z \approx \sum L(X) \Delta X$

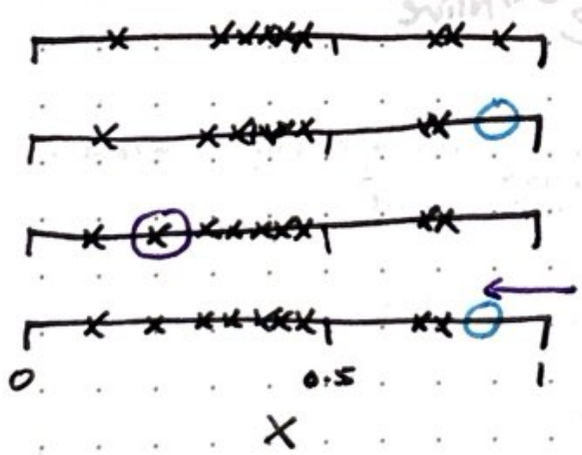


nested sampling: "combining cubes of similar likelihood & summing over them in param space" integral big integrate over contours of equal likelihood & multiplying by prior volume "under" each contour value



← nested sampling on a 2D problem

compression in 1 iterate of nested sampling:



uniformly distributed live points

remove worst

draw replacement

compression: $t \sim \mathrm{Beta}(N, 1)$

tricky bit -> can't actually see these contours, esp in high dim spaces -> so use statistical argument to estimate what they are

Estimate Compression Factor

* Don't actually know $\{X_i\}$ exactly, but can use statistical argument (to estimate it)
* At each iteration, live points are uniformly distributed between $(0, X_{i-1})$
* X_i (lowest likelihood) in current iteration, is highest X of N_{live} uniform random values

max of N_{live} values drawn from uniform dist on $[0, X_{i-1}]$ is distributed as $X_{i-1} N x^{N-1}$ therefore t_i itself is also beta distributed

* Therefore compression factor $t_i = X_i / X_{i-1}$ has beta$(N_{live}, 1)$ distribution, i.e.

$$P(t_i) = N_{live}\, t_i^{N_{live}-1}$$

highest point of these N_{live} samples has beta dist $\therefore$ t also has beta dist

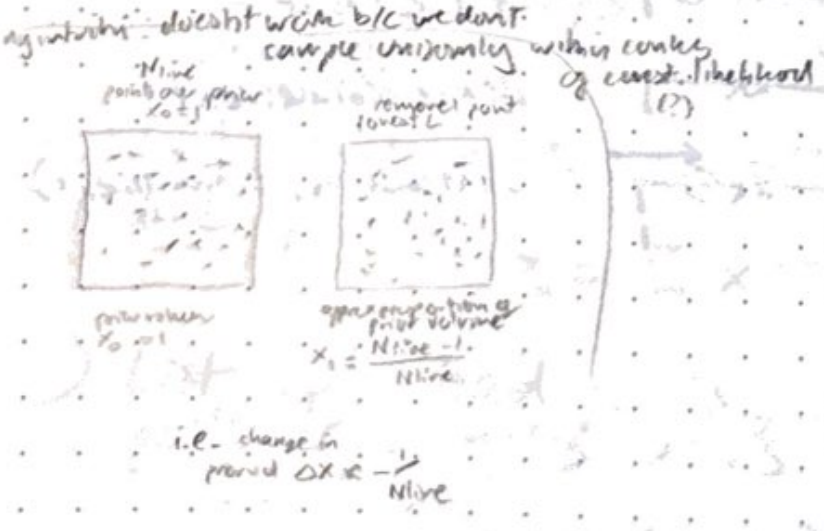
* Can approximate with mean, i.e.

$$\ln t_i \approx \langle \ln t_i \rangle = -1/N_{live}$$

or $$t_i \approx e^{-1/N_{live}}$$

$E[t] = \int_0^1 N t^{N-1}\, t\, dt = \frac{N}{N+1}$??

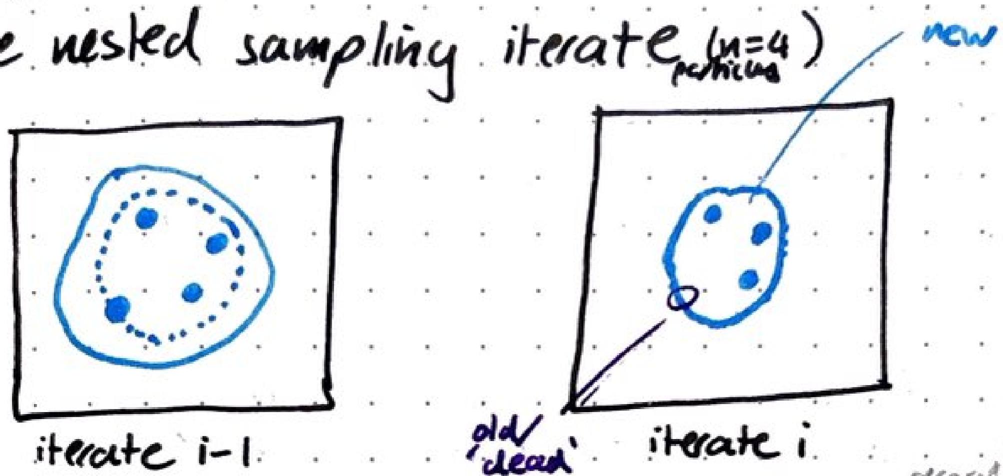
idle



Nested sampling Demo:

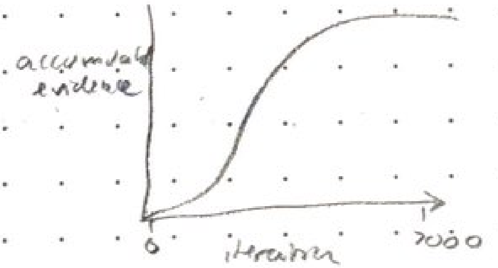
cartoon:

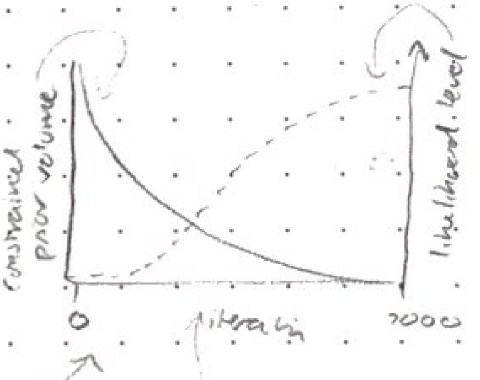
One nested sampling iterate (n=4 particles)



Demo:

- 20 gaussian uniform prior over some range e.g. -5, 5. (basically covers gaussian) → sample 1000 live points





know analytic Z = 0.01 (?) see lecture

↳ since gaussian we know analytic Z, can see nested sampling approaches correct Z (!)

(do lots of iterations likelihood of live points that remain should basically be max likelihood (x of that is ~0))

early on not accumulate much but accumulate most when prior volume & likelihood have similar values i.e. middle (whereas at either end both tend to zero so not much accumulation => S-shape curve)

(since accumulated evidence is max likelihood)

(see slides banana example, sample from banana dist. have a go if you want)

have considered basically the most basic nested sampling algorithm, but there are more!
↳ our way slows down w/ iterating → smaller & smaller region it tries to sample from

Nested Sampling: (Other Algorithms)

State-of-the-Art Nested Sampling

* Smarter ways to sample from likelihood-constrained prior, e.g. MCMC, slice sampling, sequential Monte Carlo

* Stochastically sample a set of compression factors $\{t_i\}$ rather than assume a constant mean value → gives a measure of uncertainty in evidence estimate

* Better strategic allocation of live points e.g. dynamic nested sampling

* Codes: Multinest, Ultranest, Polychord, dynesty

yay done!

useful properties of determinants:

- $\det(\underline{\underline{A}}\underline{\underline{B}}) = \det(\underline{\underline{A}})\det(\underline{\underline{B}})$ $\left(= \det(\underline{\underline{B}})\det(\underline{\underline{A}})\right)$; trivial
- $\det(\underline{\underline{A}}^T) = \det(\underline{\underline{A}})$
- $\det(\underline{\underline{A}}^{-1}) = \frac{1}{\det(\underline{\underline{A}})} = \det(\underline{\underline{A}})^{-1}$

$n \times n$ matrix $\underline{\underline{A}}$: $\det(k\underline{\underline{A}}) = k^n \det(\underline{\underline{A}})$

↳ e.g. $|2\pi\underline{\underline{\Sigma}}|^{-1/2} = (2\pi)^{-n/2}|\underline{\underline{\Sigma}}|^{-1/2}$

MCMC Theory

PREMISE:

If $\pi(a)$ satisfies D.B. : $\pi(a)\,T(b|a) = \pi(b)\,T(a|b)$, then stationary distribution π_{stat} exists! (sufficient but not necessary condition)

And additionally under the conditions 1) irreducible 2) aperiodic and 3) positive recurrent, we can guarantee that our MC will converge to the unique stationary distribution π_{stat}.

SHOW THAT OUR ALGOS ARE CONSTRUCTED S.T. DB SATISFIED:

We want to construct our algorithm s.t. DB is satisfied and $\pi_{stat}(a) = \pi_{post}(a)$, i.e. distribution we converge to is the target distribution (posterior).

$\therefore$ require $\pi_{post}(a)\,T(b|a) = \pi_{post}(b)\,T(a|b)$ ← (DB)

$$T(b|a) = J(b|a)\,A(b|a)$$

prob make the transition ↑ prob propose the jump ↑ prob accept the jump

$\Rightarrow$ we want to choose J, A s.t. they satisfy

$$\pi_{post}(a)\,J(b|a)\,A(b|a) = \pi_{post}(b)\,J(a|b)\,A(a|b)$$

$$\left(\frac{A(b|a)}{A(a|b)} = \frac{\pi_{post}(b)}{\pi_{post}(a)}\frac{J(a|b)}{J(b|a)}\right)$$

METROPOLIS SATISFIES DB: symmetric proposal dist $J(a|b) = J(b|a)$, $A(b|a) = \min\left(\frac{\pi_{post}(b)}{\pi_{post}(a)}, 1\right)$

→ suppose wlog $\pi_{post}(b) > \pi_{post}(a)$ (LHS = RHS (should really do this separately))

→ $\pi_{post}(a)\,J(b|a) = \pi_{post}(b)\,J(a|b)\,\frac{\pi_{post}(a)}{\pi_{post}(b)}$

→ since $J(b|a) = J(a|b)$, RHS = LHS!

cont from prev. page

SHOW THAT METROPOLIS-HASTINGS SATISFIES DB: (more general)

M-H: $J(a|b) \neq J(b|a)$ necessarily, $A(b|a) = \min\left(\frac{\pi_{post}(b)}{\pi_{post}(a)}\frac{J(a|b)}{J(b|a)}, 1\right)$

w.t.s. $\pi_{post}(a)\,J(b|a)\,A(b|a) = \pi_{post}(b)\,J(a|b)\,A(a|b)$

→ since either $A(b|a)=1$ or $A(a|b)=1$, easy to show LHS = RHS //

GIBBS

w.t.s. $\pi(\theta_1,\phi_1)\,T(\theta_2,\phi_2|\theta_1,\phi_1) = \pi(\theta_2,\phi_2)\,T(\theta_1,\phi_1|\theta_2,\phi_2)$

since for gibbs always accept, $T(\theta_2,\phi_2|\theta_1,\phi_1) = J(\theta_2,\phi_2|\theta_1,\phi_1) \times 1$ ($A=1$)

For gibbs start at θ_1,ϕ_1 then propose move θ then propose move ϕ (using tracked/updated conditional for both)
so $J(\theta_2,\phi_2|\theta_1,\phi_1) = \pi(\theta_2|\phi_1)\,\pi(\phi_2|\theta_2)$

~~and use fact that going $\theta_1,\phi_1 \Rightarrow \theta_1,\phi_2 \Rightarrow \theta_2,\phi_2$~~

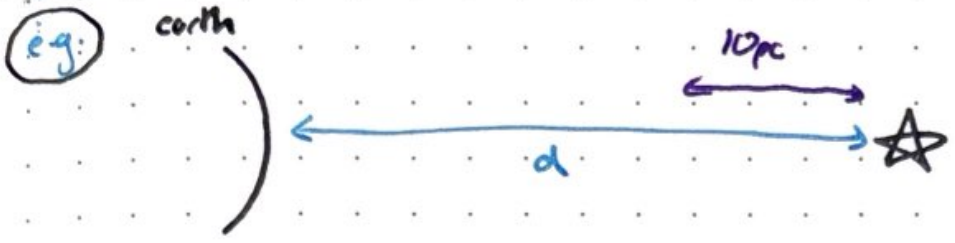
so LHS: $\pi(\theta_1,\phi_1)\,\pi(\theta_2|\phi_1)\,\pi(\phi_2|\theta_2) = \pi(\theta_1,\phi_1)\,\pi(\theta_2,\phi_2)\,\frac{\pi(\theta_2,\phi_1)}{\pi(\theta_2)\pi(\phi_1)}$

RHS: $\pi(\theta_2,\phi_2)\,\pi(\theta_1|\phi_2)\,\pi(\phi_1|\theta_1) = \pi(\theta_1,\phi_1)\,\pi(\theta_2,\phi_2)\,\frac{\pi(\theta_1,\phi_2)}{\pi(\theta_1)\pi(\phi_2)}$

start θ_1,ϕ_2, update to θ_1 as fn of ϕ_2 then update to ϕ_1 as fn of θ_1, θ_1 is indep of ϕ_2!

but since θ_1 is independent of ϕ_2 (for process on RHS)
$\pi(\theta_1)\pi(\phi_2) = \pi(\theta_1,\phi_2)$
doing similarly for LHS ⇒ LHS = RHS
⇒ DB satisfied
⇒ stationary dist $\pi_{stat} = \pi$ exists

Brief Recap on Magnitude Scale



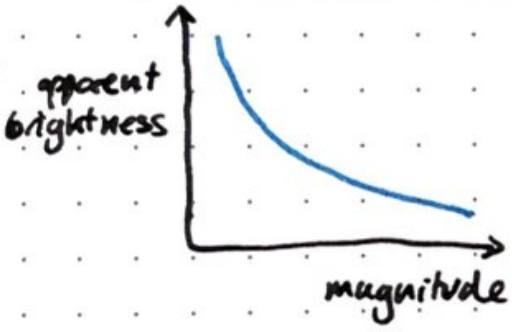
(Reminder: Flux — energy time^{-1} area^{-1} ; Luminosity — energy time^{-1})

APPARENT MAGNITUDE

define (apparent mag at distance d.)

$$m_d = -2.5 \log_{10}\left(\frac{F_d}{F_0}\right) = -2.5 \log_{10} F_d + \text{const.}$$

(m_d: apparent mag at distance d; F_d: flux at distance d; F_0: some reference flux / zero point of filter; const.: some calibration const to do with reference flux of filter)



e.g. decrease apparent magnitude by 1: star mag 5 (apparent) is $10^{4/5} \approx 2.5$ times brighter than star mag 6

lower mag = brighter

ABSOLUTE MAGNITUDE

defined as magnitude object would have if it were viewed at a distance of 10pc

$$M = m_{10pc} = -2.5 \log_{10}\left(\frac{F_{10pc}}{F_0}\right) = -2.5 \log_{10}\left(\frac{F_d}{F_0}\frac{d^2}{(10pc)^2}\right)$$

$$= -2.5 \log_{10}\left(\frac{F_d}{F_0}\right) - 5 \log_{10}\left(\frac{d}{10pc}\right)$$

(not necessarily "true" distance)

(luminosity distance: i.e. distance inferred from luminosity)

$$\boxed{M = m - 5 \log_{10}\left(\frac{d}{10pc}\right) = m - 5 \log_{10}\left(\frac{d}{pc}\right) + 5}$$

μ

Distance modulus μ def s.t. $\boxed{M = m - \mu}$ (i.e. $\mu = +5 \log_{10}\left(\frac{d}{10pc}\right)$)

(logarithmic distance) -> star at $\mu = 2$ is $\times 1.6$ times further than star at $\mu = 1$

Regression (1)

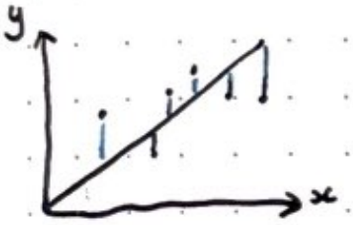
- special case of model fitting (dependent variable / independent variable)
- model describes $E[y|x]$

note: linear model = linear in params, e.g. $f(x_i;\theta) = \theta_1 + \theta_2 \sin x$
$\theta_1 + \theta_2 \sin(\theta_3 x)$ not linear model

OLS: ordinary least squares

(regress on linear model $f(x_i;\theta)$)
(assume no / negligible error in x)

e.g. simplest example: noisy data $y_i = f(x_i;\theta) + \varepsilon_i$
want to fit model $f(x_i;\theta) = \theta_0 + \theta_1 x_i$



method: find params which minimise

$$RSS = \sum_i (y_i - (\theta_0 + \theta_1 x_i))^2$$

(multiply out and find $\partial/\partial\theta_1$, $\partial/\partial\theta_0$) $\Rightarrow$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_1 = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

also, OLS estimator of variance:

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

OLS relationship to max likelihood

(why did we square residual and not e.g. take magnitude?)

assume gaussian errors in y: $P(y_i|\theta, M) = N(y_i|f(x_i;\theta), \sigma_i^2)$

Probability of observing dataset:

$$P(D|\theta, M) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}(x_i - f(x_i;\theta))^2/\sigma_i^2} = L(\theta)$$

$$\Rightarrow \ell(\theta) = -\tfrac{1}{2}\ln(2\pi\sigma_i^2) - \tfrac{1}{2}\sum_i \frac{(x_i - f(x_i;\theta))^2}{\sigma_i^2}$$

"least squares" = "max likelihood"

if data are gaussian

(2)

OLS estimate of unknown variance: $\hat{\sigma}^2 = \frac{1}{N-k}(Y-X\hat{\beta})^T(Y-X\hat{\beta})$ (k = no. of estimated parameters)

Generalise:

First introduce matrix notation $Y = X\beta + \varepsilon$

e.g. for previous example (X: N×k matrix?)

$$Y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \end{pmatrix}$$

OLS:

$$\left[RSS = (Y - X\beta)^T(Y - X\beta)\right]$$

(check using einstein summation convention notation)

(minimise wrt θ_j: $y_i^2 - 2y_i x_{ij}\theta_j + x_{ij}^2\theta_j^2$)

$$\left[\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY\right]$$

↑ unbiased!

(- weighted least squares aka χ^2 minimisation is, like OLS, another special case of generalised least squares

minimise $\chi^2 = \sum_i \frac{(y_i - x_{ij}\beta_j)^2}{\sigma_i^2}$

↳ with $W = \begin{pmatrix} \sigma_1^2 & 0 & \cdots \\ 0 & \sigma_2^2 & \\ \vdots & & \ddots \end{pmatrix}$: diagonal cov matrix)

Generalised least squares

linear model $Y = X\beta + \varepsilon$

correlated errors $W = Var(\varepsilon) = Cov(\varepsilon, \varepsilon^T)$ (known)

GLS:

$$\left[RSS = (Y-X\beta)^T W^{-1}(Y - X\beta)\right]$$

$$\left[\hat{\beta}_{GLS} = (X^TW^{-1}X)^{-1}X^TW^{-1}Y\right]$$

note: ["generalised least squares" = "max likelihood"] holds in general for gaussian errors (so for OLS & χ^2 minimisation alike!)