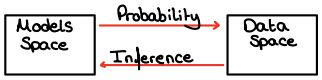
Advanced Statistical Methods for Data Science

Statistics: Collection, analysis and presentation of data. Aims to extract information from available data.

Probability: forward modelling, quantifying likelihood of events under uncertainty.

Inference Drawing conclusions from data, often probabilistic, to support/refute hypothesis



Bayesian Statistics

Probability interpreted as representing our state of knowledge / quantification of personal belief about something.

Bayes Theorem:

$$P(A,B) = P(A|B)\,P(B)$$
$$P(B,A) = P(B|A)\,P(A)$$

↳ uses Symmetry of 'and' operator (,)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

For Inference

$$P(M|D,I) = \frac{P(D|M,I)\,P(M|I)}{P(D|I)}$$

M - model (values of all parameters)

D - Data

I - All of the assumptions made during analysis

Posterior - Probability our model, M, is correct given observed Data, D and prior information, I

$P(M|D,I)$ (Bayesian only)

Likelihood - Probability of obtaining observed D data given our model M is true.

$P(D|M,I) = \mathcal{L}(D|M)$ (Frequentist as well)

Prior - Probability our model, M, is correct before any data, based on prior information, I

$P(M|I) = \pi(M)$

Evidence - total probability of observed data across all models.

$$P(D|I) = Z(D) = \int d(\text{parameters})\,\pi(\text{parameters})\,\mathcal{L}(D|\text{parameters})$$

Acts as a normalisation constant.

rewritten:

$$P(M|D) = \frac{\mathcal{L}(D|M)\,\pi(M)}{Z}$$

Likelihood: $\mathcal{L}(D|M)$

'Probability of observed data conditioned on a particular choice of the model and any model parameters'

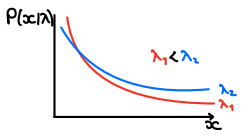
2 versions of Likelihood

Probability Distribution $\mathcal{L}(x|\theta)$ / $P(x|\theta)$: $\theta \rightarrow x$

For a given θ → function of x

Probability distribution for the data conditioned on a particular value of the models distribution.

eg: $P(x|\lambda) = \frac{\exp(-x/\lambda)}{\lambda} \rightarrow P(\underline{x}|\lambda) = \frac{\exp(-\sum_i x_i/\lambda)}{\lambda^N}$



→ Normalised with unit area

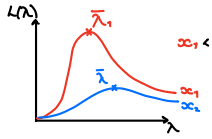
$$\int_0^\infty f(x|\lambda)\,dx = 1 \quad \forall\lambda$$

Function of model parameters $\mathcal{L}(\theta)$: $x \rightarrow \theta$

For a given x → function of θ (Inference / Inverse Problem)

$\hat{\theta}$ where $\left.\frac{\partial}{\partial\theta}\right|_{\theta=\hat{\theta}} \log(L(\theta)) = 0$

eg $L(\lambda) = \frac{\exp(-\sum_i x_i/\lambda)}{\lambda^N} \rightarrow \hat{\lambda} = \frac{\sum_i x_i}{N}$



→ Not normalised

$$\int_0^\infty f(x|\lambda)\,d\lambda \neq 1$$

Both can be considered as Slices through the Surface $\mathcal{L}(x|\theta)$

↳ For fixed θ: probability distribution of data conditioned on model parameters

For fixed x: likelihood of distribution parameters conditioned on measured result.

Fisher Information

Quantifies the information that a observation/data point, x gives about a unknown parameter, θ.

High Fisher Information ⟶ Data gives lots information on model parameter

Likelihood Dominated Regime

Score: Gradient of log-likelihood function wrt model parameter

$$S(\theta) = \frac{\partial \log \mathcal{L}(x|\theta)}{\partial \theta}$$

↳ Measures sensitivity of log-likelihood to changes in parameters

⟶ likelihood has large descriminatory power

For True parameters: θ_{true}

↳ Typically $S(\theta_{true}) \neq 0$ exactly due to statistical fluctuations in data

But: Averaged over all many samples x:

$$E_x[S(\theta_{true})] = \int dx\, S(\theta_{true})\, \mathcal{L}(x|\theta_{true})$$

($\mathcal{L}(x|\theta_{true})$: The true distribution of x for datasets)

$$= \int dx \left.\frac{\partial \log \mathcal{L}(x|\theta)}{\partial \theta}\right|_{\theta_{true}} \mathcal{L}(x|\theta_{true}) = \int \left.\frac{\partial \mathcal{L}(x|\theta)}{\partial \theta}\right|_{\theta_{true}} \frac{\mathcal{L}(x|\theta_{true})}{\mathcal{L}(x|\theta_{true})}\, dx$$

(Must be $\theta = \theta_{true}$)

$$= \int \frac{\partial \mathcal{L}(x|\theta_{true})}{\partial \theta}\, dx = \frac{\partial}{\partial \theta} \int \mathcal{L}(x|\theta_{true})\, dx = \frac{\partial}{\partial \theta} 1 = 0$$

$$\therefore \quad E_x[S(\theta_{true})] = 0$$

Fisher Information, $I(\theta)$: Variance of score at θ_{true} wrt x, $\mathrm{Var}_x(S(\theta)) \geq 0$.

$$I(\theta) = E_x[S(\theta)^2 | \theta_{true}] - E_x[S(\theta)|\theta_{true}]^2$$

$$I(\theta) = \int dx\, \mathcal{L}(x|\theta)\, S(\theta)^2 = \int dx\, \mathcal{L}(x|\theta) \left(\frac{\partial \log \mathcal{L}(x|\theta)}{\partial \theta}\right)^2$$

Peaked/Sensitive likelihood to model parameters. ⟶ Large variance of Score ⟶ Larger Fisher Information.

Alternative Definition

$$I(\theta) = -E\left[\left.\frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial \theta^2}\right|_{\theta=\theta_{true}}\right] = -\int dx\, \frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial \theta^2}\, \mathcal{L}(x|\theta)$$

Proof:

$$I(\theta) = -\int dx\, \frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial \theta^2}\, \mathcal{L}(x|\theta) = -\int dx \left\{\frac{\partial}{\partial \theta}\left[\frac{\partial}{\partial \theta} \log \mathcal{L}(x|\theta)\right]\right\} \mathcal{L}(x|\theta)$$

$$= -\int dx \left\{\frac{\partial}{\partial \theta}\left[\frac{1}{\mathcal{L}(x|\theta)} \frac{\partial}{\partial \theta} \mathcal{L}(x|\theta)\right]\right\} \mathcal{L}(x|\theta) = -\int dx \left\{\frac{1}{\mathcal{L}(x|\theta)} \frac{\partial^2}{\partial \theta} \mathcal{L}(x|\theta) - \frac{1}{\mathcal{L}(x|\theta)^2}\left(\frac{\partial}{\partial \theta} \mathcal{L}(x|\theta)\right)^2\right]\right\} \mathcal{L}(x|\theta)$$

$$= \int dx \left(\frac{1}{\mathcal{L}(x|\theta)} \frac{\partial}{\partial \theta} \mathcal{L}(x|\theta)\right)^2\right]\right\} \mathcal{L}(x|\theta) - \underbrace{\int dx\, \frac{\partial^2}{\partial \theta} \mathcal{L}(x|\theta)}_{\partial^2/\partial\theta^2 \int dx\, \mathcal{L}(x|\theta) = 0} = \int dx \left(\frac{\partial}{\partial \theta} \log \mathcal{L}(x|\theta)\right)^2 \mathcal{L}(x|\theta)$$

$$= \int dx\, S^2(\theta)\, \mathcal{L}(x|\theta)$$

QED

Transformation of Fisher Information

↳ $I(\theta)$ is not invariant under change in model parameters.

$$\phi = \phi(\theta) \longrightarrow I(\phi) = \int dx\, \mathcal{L}(x|\phi) \left(\frac{\partial}{\partial \phi} \log \mathcal{L}(x|\theta)\right)^2$$

$$\longrightarrow I(\phi(\theta)) = \int dx\, \mathcal{L}(x|\phi(\theta)) \left(\frac{\partial \theta}{\partial \phi}\right)^2 \left(\frac{\partial}{\partial \theta} \log \mathcal{L}(x|\phi(\theta))\right)^2$$

$$\boxed{I(\phi) = I(\theta)\left(\frac{\partial \theta}{\partial \phi}\right)^2}$$

Using fisher information in experiment design:

Consider taylor expansion of log likelihood:

$$\log \mathcal{L}(x|\theta) = \log \mathcal{L}(x|\theta_{true}) + (\theta - \theta_{true}) \frac{\partial \log \mathcal{L}(x|\theta)}{\partial \theta}\bigg|_{\theta=\theta_{true}} + \frac{1}{2}(\theta - \theta_{true})^2 \frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial \theta^2}\bigg|_{\theta=\theta_{true}}$$

$$E_x[\log \mathcal{L}(x|\theta)] = \underbrace{E_x[\log \mathcal{L}(x|\theta_{true})]}_{\text{Constant}} + (\theta - \theta_{true}) \underbrace{E_x[S(\theta)]}_{E[S(\theta)]=0} + \frac{1}{2}(\theta - \theta_{true})^2 \underbrace{E_x\left[\frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial \theta^2}\bigg|_{\theta=\theta_{true}}\right]}_{= -I(\theta)}$$

$$= \text{Const} - \frac{1}{2}(\theta - \theta_{true})^2 I(\theta)$$

This step cances out log: $E_x[\mathcal{L}(x|\theta)]$

$$\boxed{\exp\left[E_x[\log \mathcal{L}(x|\theta)]\right] \propto \exp\left(-\frac{1}{2}\frac{(\theta - \theta_{true})^2}{\sigma^2}\right)} \quad \text{where } \sigma = \frac{1}{\sqrt{I(\theta)}} \longrightarrow \text{High information - Narrow width}$$

↳ Not dependent on data can be evaluated without experiment

Shows likelihood behaves as a (multivariate) Gaussian about θ_{true}.

The Prior, $\pi(M)$

Describes our state of knowledge/belief about model parameters before preforming the experiment.

Typically parameters are continuous ⟶ prior is a probability distribution function

Key points/Advantages: Encodes prior knowledge/assumptions

Can penalise extreme values (regularisation)

For low data, prior dominates, guiding predictions.

Priors should accurately represent state of knowege/ignorance

↳ Informative Prior: Strong belief based on prior knowledge (large effect, dominates)

Non-Informative/Ignorance Priors: Assumes no/little prior knowledge (low effect, data dominates)

} Subject to knowledge of analyst.

Examples of Ignorance Priors

Uniform - invariant under translation

Translation Invariance:

$$\pi(x)\,dx = \pi(x + dx)\,d(x + \Delta x) \longrightarrow \boxed{\pi(x) \propto \text{Constant}}$$

Normalisation condition

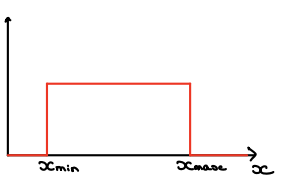
As prior is a p.d.f ⟶ must be properly normalised

↳ For uniform distribution: Requires x_{min} and x_{max} defined. [If omitted - improper prior]

If ignorant - Select value with negligible support (likelihood)

$$\boxed{\pi(x) = \frac{\mathbb{1}_{(x_{min}, x_{max})}(x)}{x_{max} - x_{min}}}$$

↳ Indicator Function: $\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$



Log Uniform - Invariant under change units/scaling. Note: Special case Jeffreys Prior

Scaling Invariance:

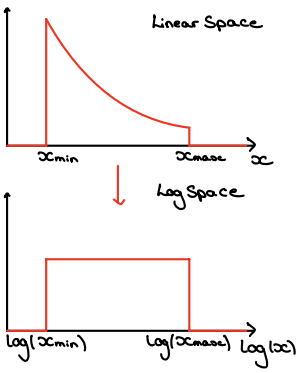
$\pi(x)\,dx = \pi(\alpha x)\,d(\alpha x) \rightarrow \boxed{\pi(x) \propto \frac{1}{x}}$

$$\boxed{\pi(x) = \frac{\mathbb{1}_{(x_{min}, x_{max})}(x)}{x \log(x_{max}/x_{min})}}$$

Comparison:

→ Rather than being uniform in linear space this uniform in log space

$\boxed{\pi(\log x) \propto \text{constant}}$ → useful for parameters that span multiple orders of magnitude.



Gives equal probability mass to each order of magnitude:

↳ ie 0.1–1, 1–10, 10–100 all have equal probability

ensure both small and large values fairly represented.

Weaknesses of Uniform:

↳ Overrepresents biases large values

For greater weighting for 100–1000 than 0.1–1 due to uniform in linear space

Jeffrey Priors

Issues with Uniform Priors when reparamaterised:

For two equivelent but equally valid parameters α and β related by invertible transform: $\beta = \phi(\alpha)$

Analysis A: Parameter α $\xrightarrow{\phi}$ **Analysis B:** Parameter $\beta = \phi(\alpha)$

Likelihood, $f(x|\alpha)$ — Likelihood $f(x|\beta) = g(x|\phi(\alpha))$

Uniform Prior, $\pi_\alpha(\alpha)$ $\xrightarrow{\text{Does not convert to uniform}}$ Induced density $|\partial\phi^{-1}/\partial\alpha|_{\phi^{-1}(\alpha)}|$ → Inconsistency → Uniform prior not Uniformative/Ignorance prior

Solution:

Jeffrey Prior on respective parameters

Square root of Fisher Information: $\pi_\alpha(\alpha) \propto \sqrt{I(\alpha)}$, $\pi_\beta(\beta) \propto \sqrt{I(\beta)}$

Advantage: Consistency under parameter transformation

'True non-informative prior'

Proof: Must conserve $\pi(\alpha)d\alpha = \pi(\beta)d\beta$

$$\pi(\beta) = \pi(\alpha)\frac{d\alpha}{d\beta} \propto \sqrt{I(\alpha)\left(\frac{d\alpha}{d\beta}\right)^2}$$

By transformation property of fisher information

$$\propto \sqrt{I(\beta)}$$

Example-Jeffry Prior Selection

For $x \in \mathbb{R}$ with likelihood $\mathcal{L}(x|N,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-N)^2}{2\sigma^2}\right)$

Jeffrey Prior for N: $\pi(N) \propto \sqrt{I(N)}$

$$I(N) = \int dx\, \mathcal{L}(x|N,\sigma)\, S(N)^2 \quad \text{where} \quad S(N) = \frac{\partial}{\partial N}\log(x|N,\sigma) = \frac{x-N}{\sigma^2}$$

$$= \int dx\, \mathcal{L}(x|N\sigma)\left(\frac{x-N}{\sigma^2}\right)^2 = \int dx\, \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-N)^2}{2\sigma^2}\right)\left(\frac{x-N}{\sigma^2}\right)^2$$

$$= \frac{1}{\sigma^2} \quad \therefore \boxed{\pi(N) \propto \sqrt{I(N)} \propto \text{Const}}$$ – For this example Jeffrey Prior, $\pi(N)$ – Uniform

Similiarly for prior on σ: $\boxed{\pi(\sigma) = \frac{1}{\sigma}}$ Jeffry Prior $\pi(\sigma)$ – log uniform.

Conjugate Priors (update knowledge without change of distribution)

A form of prior in which posterior distribution belongs to the same family as the prior after updating with data (likelihood).

$P(\theta|x) \propto \mathcal{L}(x|\theta)\,\pi(\theta)$

Posterior — Likelihood — Prior

Prior ~ Distribution A (α) $\xrightarrow{\text{Likelihood}}$ Posterior ~ Distribution (α')

The likelihood: 'updates state of knowledge using data'

For Conjugate Prior: Simply update parameters of distribution $\alpha \rightarrow \alpha'$

Unique to each Likelihood:

↳ ie. Poisson Likelihood → Gamma Prior

Proof in example

Advantages

↳ Simple form of Bayesian update (change parameters) } Useful for iterative

Allow calculations of Posterior to be Closed form } updates.

Issues:

↳ For realistic/complex likelihoods often unavailable.

Effects/Validity of varying Priors

↳ The choice of prior effects posterior and thus results

Multiple results given Subjective beliefs/knowledge

↳ Posterior does not give Single answer but 'updates State of belief'

↳ 'degree of belief' of outcome → 'willingness to bet' on outcome

Role of Prior vs Data

Low/weak data limit: Prior Dominated Regime

likelihood function is broad, does not constrain posterior

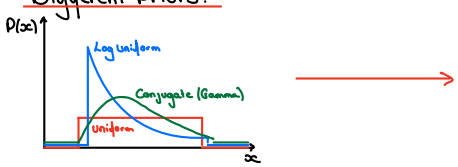
Prior has larger effect → posterior Sensitive to Subjective choice

High/Strong data limit: Likelihood Dominated Regime

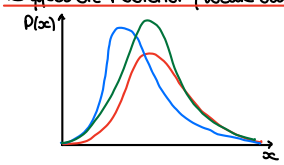
likelihood function is sharply peaked, Constraining/Concentrating posterior

Prior has small effect → posterior robust to Subjective choice

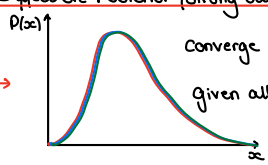
Different Priors:



Effect on Posterior (weak data)



Effect on Posterior (Strong data)



Converge to agreement given all priors (Sensible/valid).

Iterative Bayesian Inference

Update Cycle Can be preformed iteratively

Prior - represents our knowledge before date

↓

Likelihood - provides update information (applied with evidence to normalise)

↓

Posterior - represents updated knowledge after data/evidence

Iterative Process

update beliefs sequentially: Step n's posterior → Step n+1's prior

↳ Allows knowledge to accumulate

1) Initial Prior, $P(\theta)$

2) 1st Update with data, D_1 $\quad P(\theta|D_1) = \frac{P(D_1|\theta)P(\theta)}{P(D_1)}$

3) 2nd Update with data, D_2 $\quad P(\theta|D_1,D_2) = \frac{P(D_2|\theta)P(\theta|D_1)}{P(D_2)}$

↳ is this $P(D_2|\theta,D_1)$

↳ $\sum P(D_2|D_1,\theta)P(\theta|D_1)$

4) Repeating for D_3, D_4 $\quad P(\theta|D_1,D_2,D_3) = \frac{P(D_3|\theta)P(\theta|D_1,D_2)}{P(D_3)}$

$P(D_3|\theta,D_1,D_2)$

↳ $\sum P(D_3|\theta,D_1,D_2)P(\theta|D_1,D_2)$

Is equivelent?

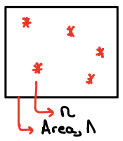
Advantages:

1) System only has to retain current posterior
   ie for large datasets where storing all data is infeasible
2) Ensures constant updating of model for data arriving sequentially (ie real world)
3) Adaptive Bayesian Updating:
   ↳ For distributions that change over time, constantly updating / tracking distribution dynamically

→ Is there a higher weighting to new evidence → Should there be for large data streams ie dilution/lagging.

Astronomy Example:

Wish to determine number density, S. Stars per square degree:



Likelihood: $\mathcal{L}(\text{Data}|\text{Model}) \propto \mathcal{L}(n|S) = \frac{(AS)^n e^{-AS}}{n!}$

($n \sim$ Poisson)

Prior choice 1: Uniform $0<S<20\ \text{deg}^{-2}$

$$\pi(S) = \frac{\mathbb{1}_{(0,20)}(S)}{20}$$

Associated Evidence $Z(n) = \int_{-\infty}^{\infty} P(n|S)\,\pi(S)\,dS$

$$= \int_{-\infty}^{\infty} \frac{(AS)^n e^{-AS}}{n!} \frac{\mathbb{1}_{(0,20)}(S)}{20}\,dS = \int_0^{20} \frac{(AS)^n e^{-AS}}{20\,n!}\,dS$$

Posterior

$$P(S|n) = \frac{(AS)^n e^{-AS}\,\mathbb{1}_{(0,20)}(S)}{20\,n!\,Z}$$

Prior Choice 2: log uniform

$$\pi(S) = \frac{\mathbb{1}_{(S_{min},S_{max})}(S)}{S\log(S_{max}/S_{min})}$$

Associated Evidence $Z(n) = \int_{-\infty}^{\infty} P(n|S)\,\pi(S)\,dS$

$$= \int \frac{(AS)^n e^{-AS}}{n!} \frac{\mathbb{1}_{(S_{min},S_{max})}(S)}{S\log(S_{max}/S_{min})}\,dS$$

Posterior:

$$P(S|n) = \frac{(AS)^n e^{-AS}}{n!\,Z} \cdot \frac{\mathbb{1}_{(S_{min},S_{max})}(S)}{S\log(S_{max}/S_{min})}$$

Prior Choice 3: Conjugate Prior

For poisson likelihood: Gamma Prior

$$\pi(S) = \begin{cases} \frac{AS^{k-1}e^{-AS/\theta}}{\Gamma(k)\,\theta^k} & \text{if } S>0 \\ 0 & 0 \end{cases}$$

Mean: $E[x] = k\theta$

Variance: $V(x) = k\theta^2$

$\Gamma(k)$ - Gamma function

Posterior

$$P(S|n) = \frac{(AS)^n e^{-AS}}{n!} \cdot \frac{AS^{k-1}e^{-AS/\theta}}{\Gamma(k)\,\theta^k} = \frac{AS^{n+k-1}e^{-\frac{AS(\theta+1)}{\theta}}}{n!\,\Gamma(k)\,\theta^k}$$

Gamma Distribution: $k' = n+k \qquad \theta' = \frac{\theta}{A\theta+1}$

Iterative Bayesian Inference Example:

Using conjugate prior for simple iteration steps.

Initial Prior $S \sim \text{Gamma}(k,\theta)$ → $k' = k+n_1,\ \theta' = \frac{\theta}{A_1\theta+1}$

Initial Posterior $S|n_1 \sim \text{Gamma}(k',\theta')$

↓ → $k'' = k'+n_2,\ \theta'' = \frac{\theta'}{A_2\theta'+1}$ → $k'' = k+n_1+n_2,\ \theta'' = \frac{\theta}{(A_1+A_2)\theta+1}$

Second Posterior $S|n_1,n_2 \sim \text{Gamma}(k'',\theta'')$

Fisher Information for Experiment Selection:

Survey 1: Area, A_1 finding n_1 stars

Survey 2: Area, A_2 ($>A_1$) finding n_2 stars (but only with probability detection $= p$)

Fisher Information Survey 1:

Likelihood $n_1|S \sim \text{Poisson}(A_1 S)$ ↳ rate

$$P(n_1|S) = \frac{(A_1 S)^{n_1} \exp(-A_1 S)}{n_1!}$$

Score $\rightarrow$ $S_1(S) = \frac{\partial}{\partial S} \log P(n_1|S) = \frac{\partial}{\partial S} \left(n_1 \log(A_1 S) - A_1 S - \log n_1!\right) = \frac{n_1}{S} - A_1$

Fisher Information

$$I_1(S) = \sum_{n_1=0}^{\infty} \mathcal{L}(n_1|S)\, S_1^2(S) = \sum_{n=0}^{\infty} \mathcal{L}(n_1|S) \left(\frac{n_1}{S} - A_1\right)^2 = \sum_{n_1=0}^{\infty} \mathcal{L}(n_1|S) \left(\frac{n_1^2}{S^2} - \frac{2 n_1 A_1}{S} + A_1^2\right)$$

↳ Discrete - Thus sum

$$= \sum_{n_1=0}^{\infty} \mathcal{L}(n_1|S) \left(\frac{n_1^2}{S^2} - \frac{2 n_1 A_1}{S} + A_1^2\right) = \frac{1}{S^2} \Sigma \mathcal{L}(n_1|S)\, n_1^2 - \frac{2A_1}{S} \Sigma \mathcal{L}(n_1|S)\, n_1 + A_1^2$$

$$= \frac{E[n_1^2]}{S^2} - \frac{2A_1}{S} E[n_1] + A_1^2 = \frac{\text{Var}(n_1) + \bar{n}_1^2}{S^2} - \frac{2A_1}{S} \bar{n}_1 + A_1^2$$

Poisson Mean: $\bar{n}_1 = A_1 S$ Variance: $A_1 S$

$$= \frac{A_1 S - A_1^2 S^2}{S^2} - \frac{2A_1}{S} A_1 S + A_1^2 = \frac{A_1}{S} + A_1^2 - 2A_1^2 + A_1^2 = \frac{A_1}{S}$$

$\rightarrow$ ie Information $\propto$ Area.

Fisher Information Survey 2:

True no. stars in A_2, m : $m|S \sim \text{Poisson}(mS)$ $\rightarrow$ m, a latent (unobserved variable)

Stars observed, n_2 : $n_2|m \sim \text{Binomial}(m, p)$

$$\mathcal{L}(n_2|S) = \sum_{m=n_2}^{\infty} P(n_2|m)\, P(m|S) = \sum_{m=n_2}^{\infty} {}^{m}C_{n_2}\, p^{n_2} (1-p)^{m-n_2} \cdot \frac{(A_2 S)^m \exp(-A_2 S)}{m!}$$

↳ $m \geq n_2$

$\nu = m - n_2$

$$= \frac{(A_2 p S)^{n_2} \exp(-A_2 S)}{n_2!} \sum_{\nu=0}^{\infty} \frac{((1-p) A_2 S)^{\nu}}{\nu!}$$

$\rightarrow \exp((1-p) A_2 S)$

$$= \frac{(A_2 p S)^{n_2} \exp(-A_2 p S)}{n_2!}$$

Itself a Poisson $\sim \text{Poisson}(A_2 p S)$

Thus using previous results: Score: $S_2(S) = \frac{n_2}{S} - A_2$ Information: $I_2(S) = \frac{A_2 p}{S}$

If $A_2 p > A_1$, $I_2 > I_1$

Survey 2 preferable

Fisher Information from Multiple Experiments:

Information adds for independant experiments.

Given independence: $\mathcal{L}(n_1, n_2|S) = \mathcal{L}(n_1|S)\, \mathcal{L}(n_2|S)$

Then: $I(S) = I_1(S) + I_2(S)$

Extracting Information from Posterior

For a high dimensional problem, posterior is complete description about state of knowledge of model parameters, θ_N

Marginal Distributions

often wish to integrate out/marginalise 'nuisance parameter

For $m+n$ parameters $\{\theta_1, \theta_2, \dots, \theta_n, \phi_1, \phi_2, \dots, \phi_m\}$:

Marginise over m parameters leaving n dim 'marginal posterior

$$P(\theta_1, \theta_2, \dots, \theta_n | x) = \int d^m\phi \, P(\theta_1, \theta_2, \dots, \theta_n, \phi_1, \phi_2, \dots, \phi_m)$$

Integrating over all parameters:

$$\int d^n\theta_N \, P(\theta_N|x) = \text{Const}$$

If correctly normalised: Const = 1

If unnormalised posterior: $P(\theta_N|x) \propto \mathcal{L}(x|\theta_N)\pi(\theta_N)$

$$\text{Const} = \text{Bayesian Evidence}, Z = \int d^n\theta_N \, P(\theta_N|x) = \int d^n\theta_N \, \mathcal{L}(x|\theta_N)\pi(\theta_N)$$

↳ Thus 'evidence' ≈ 'marginal likelihood'

Point Estimates

Summarising posterior with single/few value summary statistics.

All 'loose information' from posterior which is full information.

Posterior Mean

for θ_1:

$$\langle\theta_1\rangle = \int d^n\theta_N \, \theta_1 \, P(\theta_N|x) \quad \text{from full n-dimensional posterior}$$
$$= \int d\theta_1 \, \theta_1 \, P(\theta_1|x) \quad \text{from 1-D marginal posterior}$$

using defn of marginal

Posterior Mode - θ_N^{MAP}

Maximum a posteriori (MAP)

$$\left.\frac{\partial P(\theta_N|x)}{\partial \theta_N}\right|_{\theta_N = \theta_N^{MAP}} = 0$$

↳ Marginal posterior - integration over parameters can shift peak away from θ_N^{MAP}

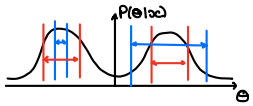
Median

$\bar{\theta}$ s.t. $\int_{-\infty}^{\bar{\theta}} d\theta_1 \, P(\theta_1|x) = \frac{1}{2}$ Only valid from 1d marginal posterior

Credible Intervals/Regions

A range/region in which model parameters lies with a particular probability level.

Issue: Not unique



Common Choices in 1D

1) Narrowest (/collection of) intervals that contain probability mass α.
   ↳ Not invariant under reparametrisation
2) Highest Density Interval
   Interval(s): $\text{Prob}(\theta \in I(f_\alpha)) \geq \alpha$ s.t. $I(f_\alpha) = \{\theta : P(\theta|x) > f_\alpha\}$ for largest f_α
3) Equal-tailed Interval:
   $\theta_{min} < \theta < \theta_{max}$: $\text{Prob}(\theta < \theta_{min}) = \text{Prob}(\theta > \theta_{max}) = \alpha/2$

Bayesian Computation

Generating Stochastic Samples from Target distribution, P (Posterior)

which can in turn help answer questions about distribution.

Monte Carlo:

Set of Stochastic Samples – n iid random variables – from Target Distribution, P(x)

$x_i \overset{iid}{\sim} P$ for $i=0,\dots,n-1$

↳ Typically sample of size 10^3–10^4

Used in Statistical Inference:

For P- posterior: → Posterior Samples → Bayesian Inference

Insights (Deterministic quantities):

1) Approximating full (d-dimensional) distribution
   ↳ d-dim histogram / Kernel density estimate
2) Approximating Marginal distributions
   ↳ Generating lower dimensional histograms.
3) Computing Credible Intervals:
   ↳ quantify uncertainties through intervals with given probabilistic content.
4) Approximate Integrals with Monte-Carlo Sum *

$\langle \varphi(x) \rangle = \int dx\, \varphi(x) P(x) \longrightarrow S_n = \frac{1}{n} \sum_{i=0}^{n-1} \varphi(x_i)$

$S_n \longrightarrow \langle \varphi(x) \rangle$ as $n \to \infty$ $\quad x_i$ - Equally weighted Stochastic Samples.

Error on estimator:

Variance $\propto 1/n$; Error $\propto 1/\sqrt{n}$ → Relatively poor preformance

↳ However holds for high dimensions → becomes best case.

Stochastic Sampling

Constructing a random Sample over Sample Space, X

given probability distribution, P.

$x \sim P, \quad x \in X$

Possible (largely) Unfeasible:

Transform Sampling:

Maps from a known distribution Q → Target Distribution, P

Given $x \sim Q$, invertible transform $y = f(x)$, $y \sim P$

$P(y) = \left| \frac{\partial f}{\partial x} \right|^{-1} P(x)$ $\quad \frac{df}{dx}$ – Jacobian of Transform

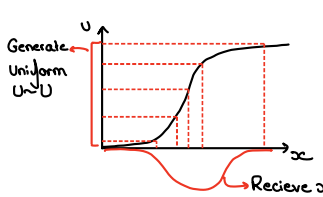
↳ 1D: Inverse CDF, → ISSUE: Many dimensions : difficult to define

Inverse CDF Method:

1) Uniform random variable $u \sim U$
2) Fuction / Transform

CDF: $F(x) = \int_{-\infty}^{x} P(x')\,dx' = u$ monotonically increasing $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$ → has unique Inverse

PPF: $F^{-1}(u) = x$ Inverse CDF, Percentage Point Function



3) $x \sim P(x)$

Rejection Sampling (accept-reject)

Proposal Distribution, Q → which we can sample from efficiently

Q must have support everywhere P does → $P(x) \leq MQ(x) \ \forall x \ 1 \leq M$

1) Random Samples from both Proposal, Q and Uniform, U
   ↳ (x, Mu) , $x \sim Q$, $U \sim U$

2) Acceptance Condition

   If $Mu < P(x) / Q(x)$: Accept

   else : reject

Iterations ≈ M iterations per sample

As Proposal Q closer to target P, $M \to 1$
↳ Efficient production Samples

Proof: $P\left(x = x \mid N < \frac{P(x)}{MQ(x)}\right) = Prob\left(N < \frac{P(x)}{MQ(x)} \mid x = x\right) \frac{Prob(x = x)}{Prob\left(N < \frac{P(x)}{MQ(x)}\right)}$

$Prob\left(N < \frac{P(x)}{MQ(x)} \mid x = x\right) \to \frac{P(x)}{MQ(x)} \cdot 1$ $\quad Prob(x = x) \to P(x)$
↳ As uniform

$Prob\left(N < \frac{P(x)}{MQ(x)}\right) \to \int_{-\infty}^{\infty} dx \frac{P(x)}{MQ(x)} Q(x) = \frac{1}{M}$
↳ as $x \sim Q$: expectation over x

Overall: $P\left(x = x \mid N < \frac{P(x)}{MQ(x)}\right) = P(x)$.

Issue: Difficulties finding proposal which
↳ Easy to sample from
Efficiently encompasses target.

Multivariate: Very large M, inefficient

Importance Sampling

↳ Determines deterministic property directly from sample of proposed distribution
↳ Does not determine samples of target distribution

Using a proposed distribution, Q close to P (meased by ϵ) – with Q support everywhere.

For a set of proposal Samples: $x_i \overset{iid}{\sim} Q \quad i = 0, 1, \ldots, n-1$

Assign Weights: $w_i = \frac{P(x_i)}{Q(x_i)}$ with weighted samples

Use of Weighted Samples: $\{(x_i, w_i) \mid i = 0, 1, \ldots, n\}$

$$I = \int dx \, \varphi(x) P(x) \approx \frac{1}{n} \sum_{i=0}^{n-1} w_i \varphi(x_i)$$

Weight contribution of $x \sim Q$
to property of P → different from rejecting

Measuring quality of proposal:
↳ best when Q is close to P → greater w_i

Number effective Samples, $n_{eff} = \frac{\left(\sum_{i=0}^{n-1} w_i\right)^2}{\sum_{i=0}^{n-1} w_i^2}$

efficiency: $\varepsilon = \frac{n_{eff}}{n}$

Markov Chain Monte Carlo

Overall aim:

1) Construct a markov chain which stationary distribution ≈ target distribution
   ↳ This will happen after chain reaches equilibrium after 'burn-in' phase.
2) Draw random samples from chain ≈ random samples from distribution

Markov Chains

↳ Overview:

For a given position/stage – transition probability, P defines probability of next move
only dependent on current position (no memory of history)

After Sufficiently evolving: Chain $\sim \pi$, Stationary distribution

With Correct Transition Probability P: $\pi \to P$

Markov Chain: Ordered Sequence of random points $x_0, x_1, \dots, x_n$ in Sample Space ($x \in \mathcal{X}$)

that Satisfy **Markov Property**:

$$P(x_{i+1} \mid x_0, \dots, x_i) = P(x_{i+1} \mid x_i) = P(x_{i+1}, x_i)$$ (Notation)

i.e only dependent on previous point in chain

↳ Transition Probability

Time-Homogenous Markov Chains: transition probability do not depend on position in chain/time: $P(x_{k+1}=x \mid x_k=y) = P(x_{j+1}=x \mid x_j=y) \quad \forall k,j$

Time Homogenous:

↳ Chain reaches Stable Stationary distributions

Used to Sample Constant target distributions.

Time Inhomogenous

↳ Chain distribution evolves dynamically

Non Stationary Underlying distribution (ie Climate Models)

Discrete Representation

For finite sample size, $\mathcal{X}$ - transition probabilities can be described by Matrice

$P_{ab}^{(k)} = P(x_{k+1}=b \mid x_k=a) \quad a,b \in \mathcal{X}$ — P - N×N matrice where $N=|\mathcal{X}|$

If inhomogeneous P depends on k

↳ Each row - Probability distribution of next update

Sum of row = 1

Disecting Prob Dist — Time Homogenous Markov Chain

For a given Starting Position:

$$P(x_2 \mid x_0) = \int dx_1 \; P(x_2 \mid x_1, x_0) \, P(x_1 \mid x_0) \quad \text{(Law Total Prob)}$$

$$= \int dx_1 \; P(x_2 \mid x_1) \, P(x_1 \mid x_0) \quad \text{(Markov Property)}$$

$$= \int dx_1 \; p(x_2, x_1) \, p(x_1, x_0) \quad \text{(Transition Probabilities)}$$

By Induction:

$$P(x_i \mid x_0) = \int dx_{i-1} \int dx_{i-2} \cdots \int dx_1 \, p(x_i, x_{i-1}) \; p(x_{i-1}, x_{i-2}) \cdots p(x_1, x_0)$$

'Distribution i^{th} point given x_0'

Deciding Transition Probability (Mappings):

Irreducibility: for any Starting point $x_0 \in \mathcal{X}$ and for any region $A \subset \mathcal{X}$:

$\exists \, n \geq 1, n \in \mathbb{Z}^+$ s.t $\int_A dx \; P(x_n \mid x_0) > 0$

↳ For any starting point, all points in space have a non zero probability of being reached, albeit not necessarily in 1 step. (go anywhere tendancy) } ergodicity

Time-Homogeneous and Irreducible

↳ Chain never Stops moving / Converge to point

But chain distribution may converge. ⟶ Limiting distribution (only Stationary dist).

Limiting Distribution: time homogenous Markov Chain approaches limiting distribution, λ on Space $\mathcal{X}$ if: $\lim_{n \to \infty} P(x_n \mid x_0) = \lambda(x_n)$ ⟶ if exists, unique

↳ If Markov Chain is ergodic: influence of x_0 dissapears over time

↳ Stationary vs Limited.

If $x_0 \sim \lambda$: Chain starts in limiting distribution (equilibrium) ⟶ If not drawn from $\lambda(x)$

x_1 from x_0: Integrate (weighted) over all possible values x_0 ⟶ Chain requires 'burn-in period' before converging

$$P(x_1) = \int dx_0 \; p(x_1 \mid x_0) \; \lambda(x_0) = \int \lambda(x_1) \, \lambda(x_0) \, dx_0 = \lambda(x_1)$$

Reality: Do not know $\lambda(x)$ — run for a bit

↳ If the Chain Starts in Stationary distribution ⟶ remains there

Equilibrium reached before collecting Samples

Stationary Distribution:

For a Markov chain with transition probabilities $p(x', x)$, a distribution on X is a stationary distribution if:

$$\pi(x') = \int dx\, \pi(x)\, p(x', x)$$ → Stay in same distribution after step

Stationary vs limited

Limited ⊂ Stationary

Stationary: Non unique - if starts in $\pi(x)$, remains there

ie Stable distribution depends on starting position, x_0

Limited: Unique - Convergent distribution of chain

ie all points converge to this

Ergodicity / Irreducibility

↳ Marko Chain has 1 Stationary distribution

Which is the limiting distribution

Non Ergodicitic / Irreducibility

↳ Cannot have limiting distribution.

Detailed Balance:

A markov chain with transition probabilities $p(x, x')$, is said to follow detailed balance wrt π if:

$$\pi(x)\, p(x', x) = \pi(x')\, p(x, x')$$

prob flux $x \to x'$ | prob flux $x' \to x$

↳ If equal - flow of probility is Symmetric

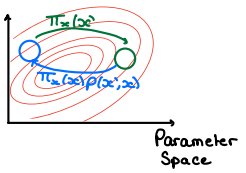
→ No net probability flux

→ Stable / Stationary

Desire $\text{prob}(x \in A \text{ and } x' \in B) = \text{prob}(x \in B \text{ and } x' \in A)$

↳ $P(x \in A) P(x' \in B \mid x)$ = ↳ $P(x \in B) P(x' \in A \mid x)$

$$\int_{x \in A} dx\, \pi(x) \int_{x' \in B} dx'\, p(x', x) = \int_{x \in B} dx\, \pi(x) \int_{x' \in A} dx'\, p(x', x)$$



Lemma #1: If Markov chain Satisfies detailed balance wrt π → π is a Stationary distribution of Markov Chain.

→ Stricter condition than neccesary.

Proof: DB condition $\pi(x)\, p(x', x) = \pi(x')\, p(x, x')$

Integrate wrt x' over X

$$\int_X dx'\, \pi(x)\, p(x', x) = \int_X dx'\, \pi(x')\, p(x, x')$$

$$\pi(x) = \underbrace{\int p(x' \mid x)\, dx'}_{1} = \int_X dx'\, \pi(x')\, p(x, x')$$

$$\pi(x) = \int_X dx'\, \pi(x')\, p(x, x')$$ → definition of Stationary. QED

Lemma #2: If Markov chain is homogenous and irreducible and Satisfies detailed balance wrt π → π is unique stationary / Limiting distribution

Result:

Aim to find transition probability $p(x', x)$ for a time-homogenous, irreducible Markov chain

St. Satisfies detailed balance where Stationary distribution, $\pi = P$, Target Distribution

result: Chain will approach P (after evolving for sufficient iterations - burn-in period)

Gibbs Sampling:

Samples are generated by iteratively updating one parameter (dimension) at a time, drawing from its 1D conditional distribution given the current values of all other parameters.

Full - High Dimensional Target Distribution: (d dimensional)

$P(x^0, x^1, \dots, x^{d-1})$

1D Conditional - for parameter x^k

$$P(x^k \mid x^0, \dots, x^{k-1}, x^{k+1}, \dots, x^{d-1}) = \frac{P(x^0, x^1, \dots, x^{d-1})}{\int dx^k \; P(x^0, x^1, \dots, x^{d-1})}$$

(All values, except variable x^k — Normalisation Integral)

Notation: $P(x^k \mid x^{k-1})$

ie $x^{k-1} = x^0, \dots, x^{k-1}, x^{k+1}, \dots, x^{d-1}$

→ assumes easily Sample from 1D conditional distribution
- Rejection Sampling
- Transform Sampling (PPF)
- Embedded/Secondary Markov Chain

Gibbs Algorithm Steps

1) $x_0 \sim \alpha$ → Arbitary Starting value chosen - drawn from initial distribution, α.
   - May be Informed - ie MLE (not importantly)

2) $i = 0$

   While $i \geq 0$ do: → Infinite iteration - formally, not in practise

   ⊣

3) $k \sim Cat(w_0, w_1, \dots, w_{d-1})$ → Decide which parameter to update by drawing from Catagorical distribution
   - $\sum_{k=0}^{d-1} w_k = 1$ → Control relative frequency of updates → Typically Uniform $- w_k = 1/d$

4) $y \sim P(x^k \mid x_i^{-k})$ → Draw random variable parameter, k from distribution Conditioned on x_i^{-k}

   $x_{i+1} = (x_i^0, \dots x^{k-1}, y, x^{k+1}, \dots, x^{d-1})$ Generate Next Sample by Subing in new 'k parameter'

Theorem: Gibbs Algorithm produces Markov Chain $x_0, x_1, \dots$ that Satisfies detailed balance condition $\pi = P$

When would you use non-Uniform Weights?

1) Assign higher weighting to Slower Converging params that mix slowly / dont vary much between Successive Iterations.

2) Some parameters posterior much more peaked than others - do not explore all values frequently
   - high Samples to correctly estimate variance.

Proof: Transition probability defined by Gibbs

$$p(y \mid x) = \sum_k w_k \; \delta^{(d-1)}(y^{-k} - x^{-k}) \; P(y^k \mid x^{-k})$$

(1-D conditional distributions)

- Sum overall possible
- Fixed / Constant - dirac delta dist

ie if $x^{-k} \neq y^{-k}$, $P = 0$

Using detailed balance eqn:

$$P(x)\,p(y \mid x) = P(y)\,p(x \mid y)$$

LHS $P(x)\,p(y \mid x) = \sum_k \frac{P(x^k \mid x^{-k}) P(x^{-k})}{P(x)} \; w_k \; \delta^{(d-1)}(y^{-k} - x^{-k}) \; P(y^k \mid x^{-k})$

→ by Dirac Delta: $P(y^{-k}) = P(x^{-k})$

RHS $P(y)\,p(y \mid x) = \sum_k P(y^k \mid y^{-k}) P(y^{-k}) \; w_k \; \delta^{(d-1)}(y^{-k} - x^{-k}) \; P(x^k \mid y^{-k})$

$$= \sum_k P(y^k \mid x^{-k}) P(x^{-k}) \; w_k \; \delta^{(d-1)}(y^{-k} - x^{-k}) \; P(x^k \mid x^{-k})$$

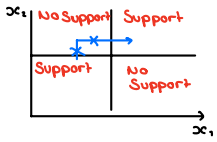
(Swap)

$$= \sum_k P(x^k \mid x^{-k}) P(x^{-k}) \; w_k \; \delta^{(d-1)}(y^{-k} - x^{-k}) \; P(y^k \mid x^{-k})$$

QED LHS = RHS.

↳ Detailed balance ensures Stationary distribution but doesn't guarentee irreducibility / limiting distribution. ie

↳ Typically produces irreducible markov chain - not guarenteed.



ie No gibbs Chain will allow transition (horizontal/vertical) between Supported regions

Inefficient Samples:

↳ by only updating 1 parameter at a time:

Common to take every $(d\text{-dim})^{th}$ point in Gibbs Chain

⟶ does not gaurentee no duplicates due to $k \sim \text{Cat}(\dots)$ but reduces them (Similiar to Sweep but not guaranteed new value cat)

Gibbs Sampling Example

2D target distribution: $P(x,y) = y\exp(-[xy+y])$ with $0 < x, y < \infty$

x Conditional: $P(x|y) = \dfrac{\exp(-yx)}{\int_0^\infty \exp(-yx)dx} = y\exp(-yx)$ (exponential Shape parameter of 1)

y Conditional: $P(y|x) \propto y\exp(-[xy+y]) \longrightarrow y|x \sim \text{Gamma}\left(k=2, \Theta=\frac{1}{x+1}\right)$ $\quad \text{Gamma}(k,\Theta) = \dfrac{y^{k-1}\exp(-y/\Theta)}{\Theta^k \Gamma(k)}$

0^{th} Iteration: $(x_0, y_0) = (1,1)$ - arbitabry

1^{st} Iteration: Randomly Choose $k \in \{0,1\}$ ⟶ Pick $k=0$, x-coordinate:

Draw $x \sim \dfrac{\exp(-y_0 x)}{y_0}$ ⟶ $(x_1, y_1) = (x_1, y_1 = y_0)$

2^{nd} Iteration: Randomly Choose $k \in \{0,1\}$ ⟶ Pick $k=1$, y-coordinate:

Draw $y \sim \text{Gamma}\left(2, \frac{1}{x_1+1}\right)$ ⟶ $(x_2, y_2) = (x_2 = x_1, y_2)$

etc:

Gibbs 'Sweep' Sampling

↳ Every Component gets updated (in a fixed order) at every iteration / in a Sweep

↳ Components updated by drawing from 1-d Conditional

Conditioned on parameter values updated so far

Does Not Satisfy detailed balance ⟶ Does Converge to target distribution.

Lemma: Gibbs Sweep algorithm produces a Markov Chain $x_0, x_1, \dots$ that has the target distribution P as unique Stationary distribution.

Algorithm:

$x_0 \sim a$

$i = 0$

While $i \geq 0$ do

for $k = 0$ to $d-1$ do:

$y^k \sim P(x^k | y^0, \dots, y^{k-1}, x_i^{k+1}, \dots, x_i^{d-1})$

↳ Not yet updated in this Sweep

↳ Already updated in this Sweep

end for

$x_{i+1} = (y^0, \dots, y^{d-1})$

$i = i+1$

Proof:

For 2d case: $P(x) = P(x^1, x^2)$ be the target distribution

Transition: $(x^0, x^1) \longrightarrow (y^0, x^1) \longrightarrow (y^0, y^1)$

$$P([y^0,y^1],[x^0,x^1]) = P(y^1|x^1)P(y^1|y^0) = \frac{P(y^0,x^1)}{\int da\, P(a,x^1)} \cdot \frac{P(y^0,y^1)}{\int db\, P(y^0,b)}$$

Defn Stationary dist: $\boxed{\pi(x') = \int dx\, \pi(x)\, p(x',x)}$

$$\int dx\, P(x)\, P(y|x) = \int dx^0 \int dx^1\, P(x^0, x^1)\, \rho([y^0, y^1], [x^0, x^1])$$

$$= \frac{P(y^0, y^1)}{\int db\, P(y_0, b)} \int dx^1 \frac{P(y^0, x^1)}{\int da\, P(a, x^1)} \int dx^0\, P(x^0, x^1)$$

$$= P(y^0, y^1) = P(y) \quad \therefore \boxed{P(y) = \int dx\, P(x)\, P(y|x)}$$ Stationary Conditioned Satisfied.

Can be extended to d dimension

Blocked Gibbs Sampling

↳ Groups multiple parameters into blocks of parameters, $\theta^N = (x^i, \ldots, x^k)$

→ allows multiple parameters (of block) to be updated at once

↳ Requires: blocks, θ^N to be chosen s.t. possible to sample from higher dimensional conditional probability. $P(\theta^N | \theta^{-N})$

Gibbs Sampling with Ordered Overrelaxation

↳ Randomly walking - Slow exploration of parameter space

OOR → Introduces negatively correlated transitions

↳ encourages chain to make larger moves

Order Statistics

↳ Draw multiple possible values from given sample

Sort them and select new value based on relative position (enforcing negative correlation)

Metropolis-Hasting Sampling

Advantage: does not require 1D conditional distribution.

The Proposal Distribution, $Q(y|x)$

A d-dimensional distribution which generates candidates for next chain position

↳ e.g commonly: $y \sim N(x, \Sigma_d)$ - Gaussian (Symmetric proposal)

Following probabilistic rule:

Proposal is either accepted ($x_{i+1} = y$) or rejected ($x_{i+1} = x_i$)

Acceptance probability:

$$a = \min\left(1, \frac{P(y)}{P(x)} \frac{Q(x|y)}{Q(y|x)}\right)$$

Algorithm:

$x_0 \sim \alpha$

$i = 0$

While $i \geq 0$:

$y \sim Q(y|x_i)$

$a = P(y)\, Q(x_i|y) / P(x_i)\, Q(y|x_i)$

$u \sim U$

If $u < a$:

$x_{i+1} = y$

else:

$x_{i+1} = x_i$

Overall transition probabilities:

↳ mixture of proposal distribution, $Q(y|x)$ with fraction a (acceptance), and delta function with fraction $(1-a)$ (rejection).

$$\rho(y, x) = \underbrace{a(y, x)}_{P(\text{accepting} | \text{proposed } y)} \underbrace{Q(y|x)}_{P(\text{proposing } y)} + \delta^d(y - x) \int dy' \underbrace{[1 - a(y', x)]}_{P(\text{proposing any } y) \cdot P(\text{rejecting} | \text{proposing } y)} Q(y'|x)$$

↳ Sum over all rejected proposals

Note: MH only depends on ratio $P(y')/P(x)$

↳ Thus does not require normalised target distribution

Do not require evidence (only $\mathcal{L}, \pi, p$)

Stuck points:

Repeat entries occur when proposal rejected ($x_{i+1} = x_i$)

↳ Acceptance fraction, g: $\frac{\text{No. accepted proposals}}{\text{Total no. iterations}}$

Metropolis Sampling:

A special case of MH where proposal is symmetric, ie:

$Q(y|x) = Q(x|y)$ → $a(y, x) = \min\left(1, \frac{P(y)}{P(x)}\right)$ The case for gaussian with identity covariance

Intuition:

'If uphill in distribution: $P(y) > P(x)$' → $a(y, x) = 1$

'If 'little downhill' in distribution: $P(y) < P(x)$ → $a(y, x) < 1$

↳ Accepted given portion of time.

'favours uphill tendency but can make downhill steps'

Theorem: MH algorithm produces Markov Chain $x_0, x_1, \dots$ that satisfies the detailed balance condition $\pi = P$

Proof:

$\pi(x)\,p(y,x) = \pi(y)\,p(x,y)$

↳ $P(x)\,a(y,x)\,Q(y|x) + \delta^d(y-x)\,P(x)\int dy'\,[1-a(y',x)]\,Q(y'|x) = P(y)\,a(x,y)\,Q(x|y) + \delta^d(x-y)\,P(y)\int dy'\,[1-a(y',y)]\,Q(y'-y)$

↳ by $\delta^{(d)}(y-x)$; $y \leftrightarrow x$ thus equivelent and cancel

Must Show: $P(x)\,a(y,x)\,Q(y|x) = P(y)\,a(x,y)\,Q(x|y)$

By definition: $a = \min\left(1, \frac{P(x)\,Q(y|x)}{P(y)\,Q(x|y)}\right)$ let $a(x,y) = \frac{P(x)\,Q(y|x)}{P(y)\,Q(x|y)}$ → $a(y,x) = 1$

↳ due to 'min' term.

Overall satisfied: [Similiarly if $a(y,x) = \frac{P(y)\,Q(x|y)}{P(x)\,Q(y|x)}$ and $a(x,y) = 1$]

Gibbs ↔ MH Algorithms

Gibbs Transition Probability $Q(y|x) = \sum_k w_k\,\delta^{(d-1)}(y^{-k}-x^{-k})\,P(y^k|x^{-k})$

MH Acceptance function: $a(y,x) = \min\left(1, \frac{P(y)\,Q(x|y)}{P(x)\,Q(y|x)}\right)$

(for Gibbs)

$$= \min\left[1,\ \frac{\sum_k w_k\,\frac{P(x^k|y^{-k})\,\delta^{(d-1)}(y^{-k}-x^{-k})}{P(x)}}{\sum_{k'} w_{k'}\,\frac{P(y^{k'}|x^{-k'})\,\delta^{(d-1)}(x^{-k'}-y^{-k'})}{P(y)}}\right]$$

$P(x) = P(x^k|x^{-k})\,P(x^{-k})$

$P(y) = P(y^k|y^{-k})\,P(y^{-k})$

$$= \min\left[1,\ \frac{\sum_k w_k\,\frac{\delta^{(d-1)}(y^{-k}-x^{-k})}{P(x^{-k})}}{\sum_{k'} w_{k'}\,\frac{\delta^{(d-1)}(x^{-k'}-y^{-k'})}{P(y^{-k'})}}\right] = 1$$

↳ Always unity / always accepted.

Gibbs — Special case of MH where one uses specific proposals built from the 1D conditionals of the target

- With an acceptance of unity.

Limitations of Gibbs and MH

↳ Exhibit random walk behaviour → follows diffusion type behaviour

May take many iterations to 'diffuse' slowly around sample space

↳ **Diffusion Equator**

'Root Mean Squared' Distance moved $\sim \sqrt{N_{steps}}$

'Stuck' in Extreme regions

Gibbs – Large No. Small Steps due to 1d conditions

MH – Large Proposal dist – Large No. rejections

Small Proposal dist – Small Steps

↳ especially bad for high dimensions

Hamiltonian/Hybrid Monte Carlo

↳ Exploits gradient information about the target distribution ($\nabla_x \log P$) → Treats distribution as potential well.

Proposes a transition to a new position y by drawing a random momentum $p \sim Q$ and integrating Hamiltonians equations forward.

✓: only requires a function proportional to the target PDF. (not normalised)

X: require gradient of Target distribution wrt parameters.

↳ $-\log P(x)$ wrt x: Can only be applied to smooth target distributions.

Hamiltonian Dynamics

Positions - Points $x \in X$ in Sample Space

Potential / Log Position Distribution (Exploit - Logarithm of target PDF)

$$\log P(x) = -E(x) + \text{const}$$

↳ $E(x)$ – potential

Force (acting on a particle)

$$\text{Force} = -\nabla E(x)$$

Momenta -

Variables ($\rho \in T_x X$) in tangent Space. – Same dimensions as X

Kinetic / Log Momenta Distribution, Q on Tangent Space, $T_x X$

Can be chosen freely but easy to sample from - ie Multivariate gaussian

$$\log Q(\rho) = -K(\rho) + \text{const} \longrightarrow \text{Kinetic Energy, } K(\rho) \qquad K(\rho) = \frac{1}{2} \rho^T M^{-1} \rho$$

↳ M - mass matrix, positive Semidefinite

Phase Space, (x, ρ)

Position, momentum pairs - dimensionality: x2 original X

Canonical Distribution, R

distribution on phase space (-ve of Hamiltonian)

$$\log R(x,\rho) = -H(x,\rho) + \text{const} \longrightarrow \text{Hamiltonian: } H(x,\rho) = E(x) + K(\rho)$$

$$H(x,\rho) = -(\log Q(\rho) + \log P(x)) + \text{const}'$$

$$= -(\log\{Q(\rho) P(x)\}) + \text{cost}'$$

Sampling

↳ Sample $(x, \rho) \sim R$ in 2d dimensions (rather than $x \sim P$)

Discard Momentumn, ρ variable

Equivalence as: $P(x) = \int d\rho \, R(x,\rho)$

Time parameters: (introduced parameters)

Position and Momentum promoted to function of $t \rightarrow x(t), p(t)$

$x(t), p(t)$ evolve by Hamiltonian Dynamics:

$$\frac{dx^k}{dt} = \frac{\partial H}{\partial p^k} = (\underline{M}^{-1} \cdot \underline{p})^k$$

$$\frac{dp^k}{dt} = -\frac{\partial H}{\partial x^k} = -\nabla E(x)$$

Deterministic Map, T_s

$T_s : (x(t), p(t)) \longrightarrow (x(t+s), p(t+s))$

Properties of Hamiltonian Dynamics

↳ **Reversibility**

→ Hamiltonian dynamics is reversible: T_s has an inverse T_{-s}

$T_s(x(t), p(t)) \longrightarrow x(t+s), p(t+s)$ } reverse momenta, get back to the same

$T_{-s}(x(t+s), -p(t+s)) \longrightarrow x(t), -p(t)$ } initial point with same momenta (in reverse)

Volume Preservation / Liouville's Theorem

→ Consider all points (x, ρ) in region of Phase Space, A (volume V_A)

B is the image of A under T_s; the its volume $V_B = V_A$

Hamiltonian Conservation

Hamiltonian dynamics Conserved Hamiltonian ($\partial H/\partial t = 0$)

Numerical Integration/Discrete Steps

↳ exact Solutions often not feasible

Must use Symplectic Integrator → Exactly time reversible
Volume Preserving.

Commonly: Leapfrog Method (Symplectic)

↳ 1) Half Step in momentum: $p_i(t+\frac{1}{2}\Delta t) = p(t) + \frac{1}{2}\Delta t \frac{dp}{dt}$

$= p(t) - \frac{1}{2}\Delta t \nabla E\big|_{x=x(t)}$ → Energy term evaluated at initial position, x_t

2) Full Step in Position: $x(t+\Delta t) = x(t) + \Delta t \frac{dx}{dt}$

$= x(t) + \Delta t\,(M^{-1} \cdot p(t+\frac{1}{2}\Delta t))$ → Momentum term evaluated at half Step

3) Half Step in momentum: $p_i(t+\Delta t) = p(t+\frac{1}{2}\Delta t) + \frac{1}{2}\Delta t \frac{dp}{dt}$

$= p(t+\frac{1}{2}\Delta t) - \frac{1}{2}\Delta t \nabla E\big|_{x=x(t+\Delta t)}$ → Energy term evaluated at ½ Step position, $x(t+\frac{1}{2}\Delta t)$

Properties of Leapfrog Dynamics (match that of Hamiltonian).

↳ Reversibility.

Leapfrog dynamics are exactly reversible — for all Δt

Forwards:

$x(t), p(t) \longrightarrow x(t+\Delta t), p(t+\Delta t)$

1) $p(t+\Delta t/2) = p(t) - \frac{\Delta t}{2}\nabla E(x(t))$

2) $x(t+\Delta t) = x(t) + \Delta t\, M^{-1} p\left(t+\frac{\Delta t}{2}\right)$

3) $p(t+\Delta t) = p\left(t+\frac{\Delta t}{2}\right) - \frac{\Delta t}{2}\nabla E(x(t+\Delta t))$

Reverse – Reverse Momentum $p(t+\Delta t) \longrightarrow -p(t+\Delta t)$

1) $p'(t+\frac{\Delta t}{2}) = -p(t+\Delta t) - \frac{\Delta t}{2}\nabla E(x(t+\Delta t))$

$= -\left[p\left(t+\frac{\Delta t}{2}\right) - \frac{\Delta t}{2}\nabla E(x(t+\Delta t))\right] - \frac{\Delta t}{2}\nabla E(x(t+\Delta t))$

$= -p\left(t+\frac{\Delta t}{2}\right)$

2) $x'(t) = x(t+\Delta t) + \Delta t M^{-1} p'(t+\Delta/2)$

$= x(t) + \Delta t\, M^{-1} p\left(t+\frac{\Delta t}{2}\right) + \Delta t M^{-1}\left(-p\left(t+\frac{\Delta t}{2}\right)\right) = x(t)$

3) $p'(t) = p'\left(t+\frac{\Delta t}{2}\right) - \frac{\Delta t}{2}\nabla E(x'(t))$

$= -p\left(t+\frac{\Delta t}{2}\right) - \frac{\Delta t}{2}\nabla E(x(t)) = -p(t) + \frac{\Delta t}{2}\nabla E(x(t)) - \frac{\Delta t}{2}\nabla E(x(t))$

$= -p(t)$

$x(t+\Delta t), -p(t+\Delta t) \longrightarrow x(t), -p(t)$ QED.

Volume preserving (exactly)

Note all three Steps are 'Shear transformations':

↳ each variables (x or p) changes by an amount that depends linearly on the other

eg Step 2: $\begin{pmatrix} x \\ p \end{pmatrix} \longrightarrow \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x + \Delta t f(p) \\ p \end{pmatrix}$ with Jacobian $J = \begin{vmatrix} \partial x'/\partial x & \partial x'/\partial p \\ \partial p'/\partial x & \partial p'/\partial p \end{vmatrix} = \begin{vmatrix} 1 & \Delta t\, \partial f/\partial p \\ 0 & 1 \end{vmatrix} = 1$

↳ All Steps – Unit determinant. Thus overall volume preserving

Hamiltonian Conservation (Approximately for Small Δt)

Not perfect Conservation due to errors from discrete Steps.

$H(x(t+\Delta t), p(t+\Delta t)) = H(x(t), p(t)) + \mathcal{O}(\Delta t^2)$

Algorithm

$x_0 \sim \alpha$

$i = 0$

While $i \geq 0$:

$p \sim Q$ → Draw momentum from distribution Q

$x = x_i$

$H_{initial} = H(x,p)$ → Initial energy

for $l=1$ to L do:

$x, p \longrightarrow Leapfrog(x, p, \Delta t, M)$ → Propogate Forwards with Leapfrog

$H_{final} = H(x,p)$ → Final energy

$a = \exp(H_{initial} - H_{final})$ → Measure of change in Hamiltonian (1 if equal)

Measure of 'failure' of discrete integration to conserve H.

$U \sim N$ → Draw random variable from uniform $[0,1]$

if $U < a$: → MH accept/reject process:

$x_{i+1} = x$ — Accept with probability $\min(1, a)$

else

$x_{i+1} = x_i$

$a > 1$: $H_{initial} \geq H_{final}$ → 100% acceptance

$a < 1$: $H_{initial} < H_{final}$ → <100% (but close) acceptance

Increase in H is unfavourable

$H(t) \approx H(t+\Delta t)$ to $\mathcal{O}(t^2)$

HMC ⟷ Blocked Gibbs + Metropolis Hastings

With Augmented distribution: $R(x,p) = P(x)Q(p)$

↳ R is Seperable: $P(x), Q(p)$ – Marginal and Conditional distribution $P(x) = R(x|p)$

$Q(p) = R(p|x)$

Advantages:

↳ For large L, Small Δt:

Large Steps with high acceptance probability

When Sampling $(x_i, p_i) \overset{iid}{\sim} R$

↳ Blocked Gibbs algorithm: Block 1: position variables

Block 2: Momentum variables

$p_{i+1} \sim R(p|x_i) \equiv Q(p)$ → Easy to sample from. (Q - often gaussian)

$x_{i+1} \sim R(x|p_{i+1}) \equiv P(x)$

↳ MH Step: Proposal of position: Leapfrog with L steps (integrate)

↳ Acceptance: $\min(1, a)$

Theorem: HMC algorithm produces Markov Chain x_0, x_1, x_2 that satisfies detailed balance with $\pi = p$

Proof: Shown by proving HMC special case of Metropolis.

i) Suppose initial position, x

ii) Momentum - Draw random p and evolve (numerically integrate) forwards - Leapfrog

↳ Deterministic Integration (ie fixed output for given input)

∴ HMC provides a proposal distribution.

$y, q = I(x,p)$

$\text{Proposal}(y|x) = \frac{Q(p)}{|\partial y / \partial p|}$ → dependant on Probability (p) from dist Q → where is this from

∴ By exact reversibility of leapfrog Hamiltonian Dynamics.

$x, p = I(y, -q)$

$\text{Proposal}(x|y) = \frac{Q(-q)}{|\partial x / \partial q|}$

By Leapfrog hamiltonian dynamics: Volume preserving

$\frac{\partial y}{\partial p} = \frac{\partial x}{\partial q}$

∴ given $Q(p) = Q(-p)$

$$\frac{\text{Proposal}(x|y)}{\text{Proposal}(y|x)} = \frac{Q(p)}{Q(q)}$$

Thus acceptance probability

$$a(y,x) = \min\left(1, \frac{P(y)\,\text{Proposal}(x|y)}{P(x)\,\text{Proposal}(y|x)}\right) = \min\left(1, \frac{P(y)Q(q)}{P(x)Q(p)}\right) = \min\left(1, \frac{\exp(-H(y,q))}{\exp(-H(x,p))}\right)$$

$$a(y,x) = \min\left(1, \exp\left[H(x,p) - H(y,q)\right]\right) = \min\left(1, \exp\left[H_{initial} - H_{final}\right]\right)$$

↳ even for large proposal steps ⟶ High acceptance probability

$a = \min\left(1, \exp(H_{initial} - H_{final})\right)$ Still close to 1 as discretised Hamiltonian dynamics approximately conserves Hamiltonian

Hyperparameters of HMC:

↳ time Step Δt ⟶ if too large, Leapfrog Integration inaccurate (H not conserved − a very low)

Number of integrations, L ⟶ if too large, unnecessary integration − particle oscillates about potential well

↳ if $L\Delta t$ too small - does not explore space quickly.

Mass Matrix, M (often Identity matrix)

No U-Turn Sampler:

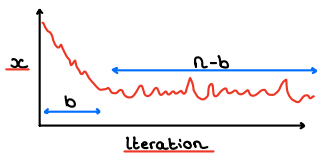
↳ Solves problem of tuning L by adaptively stopping Hamiltonian trajectory when it double backs on its self. $(q_{new} - q_{old}) \cdot q_{new} < 0$

↳ Once halted, point selected by weighted probabilities randomly.

⟶ avoids unnecessary computation while maintaining exploration efficiency

Convergence Diagnostics

Burn in period:



b-burn in: 'Initial transient part of the Markov Chains evolution - remembers starting position. Discarded as not representative of chain / target distribution

Gelman-Rubin Statistic

Checks for consistency between intra and inter chain estimates of the variance of an arbitary function, $f(x)$.

Intra-Chain Variance: Variance within each individual MCMC chain.

Inter-Chain Variance: Variance between multiple independent chains.

For N_c chains of length n (post burn-in)

x_{ij} − i denotes position $i = 0, 1, \ldots, n-1$

− j denotes chain $j = 0, 1, \ldots, N_c - 1$

GR Statistic: $R_f = \sqrt{\frac{B}{nW}}$ ⟶ $1/n$ allows Comparibility − As B- Variance of chains of length $n \times N_c$

W- Variance of chains of length N_c

B-Between-Chain (inter chain) Variances (Variance of Chain means $\times n$ − Variance of all Samples)

$$B = \frac{n}{N_c - 1} \sum_{j=0}^{N_c-1} \left(E[f]_j - \bar{f}\right)^2$$ ⟶ Why $\times n$: estimating variance underlying distribution rather than means:

Corrects for $\text{Var}(\bar{x}) = \sigma^2/n$

↳ Mean of j^{th} chain: $E[f]_j = \frac{1}{n}\sum_{i=0}^{n-1} f(x_{ij})$ Overall Sample Mean: $\bar{f} = \frac{1}{N_c}\sum_{j=0}^{N_c-1} E[f]_j$

W - Within-Chain (intra chain) variances (Mean of Chain variances)

$$W = \frac{1}{N_c}\sum_{j=0}^{N_c-1} \mathrm{Var}[f]_j$$

↳ Variances of j^{th} chain: $\mathrm{Var}[f]_j = \frac{1}{n-1}\sum_{i=0}^{n-1}(f(x_{ij}) - E[f]_j)^2$

Evaluating R_f:

R_f - depends on function f

↳ Typically evaluated on individual parameters: $f(x) = x^k$ — d-dimensional

R_k - Convergence param for k^{th} param. $k = 0, 1, \dots, d-1$ — Parameter Space

↳ Parameters converge/explore at different rates.

R_f ideally 1: Unreasonable

$R_{f,k} \leq 1.2 \quad \forall k$ (dimensions) ⟶ expected deviation due to Stochasity.

Autocorrelation Length:

Once a MCMC chain have reached equilibrium — wish to draw independant Samples from it.

Adjacent Samples in MCMC Chain - not independant. (each correlated with predecessor)

Thinning - Take every m Samples

↳ Measure of Correlation:

Autocorrelation Length - Integrated Autocorrelation Time (IAT), τ_f

↳ Length of time for chain Samples to be approximately independent.

Effective Samples from n chain items:

↳ after burn out, b and thinning, m

$$N_{eff} = \frac{n-b}{m}$$

Expectation $f(x)$: $\bar{f} = \langle f \rangle = \int f(x) P(x)\,dx = \frac{1}{n}\sum_{i=0}^{n} f(x_i)$

If Chain were independent: $x_i \sim iid$

$\mathrm{Var}[S_n] = \frac{1}{n}\sigma_f^2$ $\quad \sigma_f^2 = \int dx\,(f(x) - \langle f \rangle)^2 P(x)$

↳ however variance larger by τ_f

In fact:

Autocorrelation function at lag I:

$\rho_f(I) = \frac{C(I)}{C(0)}$ where $C(I) = \frac{1}{n-I}\sum_{i=0}^{n-I-1}(f(x_i) - \bar{f})(f(x_{i+I}) - \bar{f})$

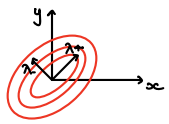
$\tau_f = 1 + 2\sum_{I=0}^{n-1}\rho_f(I)$ typically calculated for each parameter, τ_k

IAT: $\tau = \max_k(\tau_k)$

For safety often $m = 10\tau$

Case Study: Strongly Correlated ($\beta = 0.9$) two-Dimensional Gaussian

$$P = N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \beta\\ \beta & 1\end{pmatrix}\right) \propto \exp\left(-\frac{1}{2}(x-N)\Sigma^{-1}(x-N)^T\right) = \exp\left(\frac{-[x^2+y^2-2\beta xy]}{2[1-\beta^2]}\right)$$



$$\frac{\lambda_+}{\lambda_-} = \frac{1+\beta}{1-\beta} \approx 19$$

Gibbs Sampling

1-D conditional distributions:

$P(x|y) \propto \exp\left(\frac{-(x-\beta y)^2}{2[1-\beta^2]}\right)$; $x|y \sim N(\beta y, 1-\beta^2)$

$P(y|x) \propto \exp\left(\frac{-(y-\beta x)^2}{2[1-\beta^2]}\right)$; $y|x \sim N(\beta x, 1-\beta^2)$

Improvement: Define new rotated variables

$(u, v) = R_{\pi/4}(x, y)$

Metroplis-Hasting Sampling (Metropolis)

Preposal Distribution

$Q = N(\underline{0}, \underline{\Sigma})$ i) $\underline{\Sigma}_{Small} = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$ $\underline{\Sigma}_{large} = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}$ $\underline{\Sigma}_{Corr} = \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}$

↳ Effictively using Samples from P to produce Samples from P.

MH Sampling: Weighing up:

Too Small a preposal: High acceptance rate
Slow random walk behaviour

Too large a preposal: Low acceptance rate
bigger Steps.

} Balance by minimising IAT, τ

Hamiltonian

Requires partial derivative of the energy function:

$\log P(x) = -E(x) + \text{const}$

$$E(x,y) = \frac{x^2 + y^2 - 2\beta xy}{2[1-\beta^2]} \longrightarrow \frac{\partial E}{\partial x} = \frac{x - \beta y}{1-\beta^2}, \quad \frac{\partial E}{\partial y} = \frac{y - \beta x}{1-\beta^2}$$

Free Parameters: M, δt, L

↳ Optimised to allow large Steps in target distribution with high acceptance rate.

Preformance Metrics Table

Algorithm	IAT, τ	Acceptance Frac, g	No. ud Samples, Neff (%)	time/iteration	time/Sample
⋮	⋮	⋮	⋮	⋮	⋮

Bayesian Model Selection

↳ Comparison of models

Pre pose 2 models:

	Prior	Likelihood.
Model A: λ_A	$\pi(\lambda_A \mid A)$	$\mathcal{L}(D \mid \lambda_A, A)$
Model B: λ_B	$\pi(\lambda_B \mid B)$	$\mathcal{L}(D \mid \lambda_B, B)$

Biases in MLE Ratio:

$$\text{MLE Ratio} = \frac{\max_{\lambda_A} \mathcal{L}(D \mid \lambda_A, A)}{\max_{\lambda_B} \mathcal{L}(D \mid \lambda_B, B)} = \frac{\mathcal{L}(D \mid \hat{\lambda}_A, A)}{\mathcal{L}(D \mid \hat{\lambda}_B, B)}$$

→ MLE Ratio { >1: Model A 'better' ; <1: Model B 'better' }

↳ Issues:

↳ Fails to account for complexity of model

A model with many parameters → generally able to fit data better → Higher value of MLE

Suceptible to prioritising overfitting through more complex model.

[Parsimonious - model that fits data well with Small no. parameters]

The Odds Ratio

↳ Alternatively Bayesian Approach : Posterior Odds Ratio

$$\text{Posterior Odds Ratio, } \mathcal{O}_{A,B} = \frac{P(A|D)}{P(B|D)} = \frac{P(D|A) \times P(A)}{P(D|B) \times P(B)}$$

$$\mathcal{O}_{A,B} = \frac{1}{\mathcal{O}_{B,A}}$$

↳ Prior Odds Ratio – Encompasses prior belief of models (A-priori)

1, Unity for fair/unbiased comparison

Marginalise over Parameters

$$\mathcal{O}_{A,B} = \frac{\int d\lambda_A \, P(D,\lambda_A|A) \times P(A)}{\int d\lambda_B \, P(D,\lambda_B|A) \times P(B)}$$

Product rule – Achieves Likelihood × Prior

$$\mathcal{O}_{A,B} = \frac{\int d\lambda_A \, P(D|\lambda_A,A) \, P(\lambda_A|A) \times P(A)}{\int d\lambda_B \, P(D|\lambda_B,A) \, P(\lambda_B|B) \times P(B)} = \frac{\int d\lambda_A \, \mathcal{L}(D|\lambda_A,A) \, \pi(\lambda_A|A) \times P(A)}{\int d\lambda_B \, \mathcal{L}(D|\lambda_B,A) \, \pi(\lambda_B|B) \times P(B)}$$

Note: $\int d\lambda_A \, \mathcal{L}(D|\lambda_A,A) \, \pi(\lambda_A|A) = Z_A$

$\int d\lambda_B \, \mathcal{L}(D|\lambda_B,A) \, \pi(\lambda_B|B) = Z_B$

$$\mathcal{O}_{A,B} = \frac{Z_A}{Z_B} \times \frac{P(A)}{P(B)}$$

↳ Ratio of model evidences updates our relative state of beliefs of two models

Influence of Prior Odds Ratio.

↳ Prior odds ratio generally insignificant in comparison to evidence ratio

↳ i.e. If model fits data significant differently – evidence ratio has greater influence

If both models preform similiarly (Evidence Ratio ≈ 1) or increase does not justify Complexity:

↳ Prior belief dominates.

Example:

Model A: $P(D|A) = \prod_{i=1}^{6} \frac{\exp(-\frac{1}{2}y_i^2)}{\sqrt{2\pi}} = 1.41 \times 10^{-4}$ → Gaussian Mean: 0, Variance: 1

(no free parameters – no marginalisation integral)

Max MLE/Mean MLE $= 1.41 \times 10^{-4}$ Evidence = Likelihood

Model B: $P(D|\lambda,B) = \prod_{i=1}^{6} \frac{\exp(-\frac{1}{2}(y_i-\lambda)^2)}{\sqrt{2\pi}}$ → Gaussian Mean: λ

(λ - only free parameter)

Max MLE ($\lambda = \hat{\lambda} = 1$) $= 2.84 \times 10^{-3}$

↳ Calculate Evidence:

Prior: $P(\lambda|B) = \frac{1}{2\Lambda} \mathbb{1}_{(-\Lambda,\Lambda)}(\lambda)$ where $\Lambda \gg 1$.

$$P(B|D) = Z_B = \int_{-\Lambda}^{\Lambda} d\lambda \, \frac{1}{2\Lambda} \prod_{i=1}^{6} \left[\frac{\exp(-\frac{1}{2}(y_i-\lambda)^2)}{\sqrt{2\pi}} \right]$$

Limit of a wide, uninformative prior, $\Lambda \to \infty$:

$$P(B|D) \approx \frac{1}{2\Lambda} \int_{-\infty}^{\infty} d\lambda \, \prod_{i=1}^{6} \left[\frac{\exp(-\frac{1}{2}(y_i-\lambda)^2)}{\sqrt{2\pi}} \right] = \frac{1.45 \times 10^{-3}}{\Lambda}$$

Posterior Odds Ratio:

$$\mathcal{O}_{A,B} = \frac{P(A|D)}{P(B|D)} \times \frac{P(A)}{P(B)} = \frac{1.41 \times 10^{-4}}{1.45 \times 10^{-3}/\Lambda} \times 1 = 0.0973\,\Lambda$$

↳ Effected by both prior odds ratio and model priors.

↳ Penalises both large no. parameters and weakly constrained a-piriori

Computing the Bayesian evidence

Bayes Theorem: $P(x|d) = \frac{P(d|x,M)\,P(x|M)}{P(d|M)}$

Bayesian Evidence: $\boxed{Z = P(d|M) = \int_x dx\, \mathcal{L}(d|x)\,\pi(x)}$

→ Integration over full space of model parameters, $x \in \mathcal{X}$

Methods for Computation:

In general - Numerically evaluating a complicated/high dimensional integral over the full parameter space

→ **Analytic Computation**

↳ Often only feasible for well defined - low dimensional posterior. (often a conjugate prior)

Laplace Approximation.

↳ **Unnormalised Posterior Distribution** - $\hat{P}(\vec{x}) = \hat{P}(D|\vec{x}) = \mathcal{L}(\vec{x})\,\pi(\vec{x})$

Normalising Evidence - $Z = \hat{P}(D) = \int d\vec{x}\, \hat{P}(\vec{x})$

Suppose $\hat{P}(\vec{x})$ has maximum at: $\hat{\vec{x}}$

Expand $\log \hat{P}(\vec{x})$ about $\hat{\vec{x}}$ (general case: multivariate D-dimensional problem)

$$\log \hat{P}(\vec{x}) \approx \log \hat{P}(\hat{\vec{x}}) - \frac{1}{2} C_{ij} (x_i - \hat{x}_i)(x_j - \hat{x}_j) + \mathcal{O}([\vec{x}-\hat{\vec{x}}]^3)$$

↳ No linear term as expanded about maximum

↳ Ignoring higher order terms.

C - Hessian Matrix

↳ Negative matrix of second derivatives of log-posterior:

$$C_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j}\bigg|_{\vec{x}=\hat{\vec{x}}} \log \hat{P}(\vec{x})$$

Exponential:

$$\hat{P}(x) \approx \hat{P}(\hat{x}) \exp\left(-\frac{c}{2}(x-\hat{x})^2\right)$$

↓ Unnormalised Posterior ≈ Gaussian (well defined numerical integral)

$$Z \approx \hat{P}(\hat{x}) \sqrt{\frac{(2\pi)^D}{\det C}}$$

→ Semi Analytic Method

↳ Max. $\hat{\vec{x}}$ and C_{ij}'s derivatives must be found.

Note: Not invariant under change of parameters

$x \longrightarrow x' = y(x)$ then $\boxed{C \longrightarrow C' = \frac{C}{(dy/dx)^2}}$

Proof: (1-D case) $C = -\frac{d^2}{dx^2}\big|_{x=\hat{x}} \log P(x)$ $\quad C' = -\frac{d^2}{dy^2}\big|_{y=\hat{y}} \log P(y)$

$$\frac{d^2}{dx^2} = \frac{d}{dx}\left(\frac{d}{dy}\frac{dy}{dx}\right) = \frac{d^2y}{dx^2}\cdot\frac{d}{dy} + \frac{dy}{dx}\frac{d}{dx}\left(\frac{d}{dy}\right) = \frac{d^2y}{dx^2}\cdot\frac{d}{dy} + \frac{dy}{dx}\left(\frac{d}{dy}\frac{dy}{dx}\right)\frac{d}{dy} = \frac{d^2y}{dx^2}\cdot\frac{d}{dy} + \left(\frac{dy}{dx}\right)^2\frac{d^2}{dy^2}$$

$$\underbrace{\frac{d^2}{dx^2}\log P(x)}_{-C} = \frac{d^2y}{dx^2}\underbrace{\frac{d}{dy}\log P(y(x))}_{\frac{d}{dy}\log(P(y))|_{y=\hat{y}}=0} + \left(\frac{dy}{dx}\right)^2\frac{d^2}{dy^2}\log P(y(x)) = \left(\frac{dy}{dx}\right)^2\underbrace{\frac{d^2}{dy^2}\log P(y(x))}_{-C'}$$

$$C = \left(\frac{dy}{dx}\right)^2 C' \longrightarrow \boxed{C' = \frac{C}{(dy/dx)^2}}$$

Savage-Dickey Density Ratio

↳ Allows calculation of evidence ratio (Bayes factor) (not Z explicitly) for nested models:

i.e. $B_{1,2} = \frac{P(d|M_1)}{P(d|M_2)} = \frac{Z_1}{Z_2}$

Requires: i) 1 model to be nested inside other

ii) Consistent priors

Useful for: Bayes factor allows for model comparison - quantifies a model's ability to explain observed data over another

$B_{1,2} > 1$: Data favours M_1

$B_{1,2} < 1$: Data favours M_2

Nested Models:

M_1: $\emptyset$ (lower dimensionality)

M_2: $\varepsilon, \emptyset$ (higher dimensionality)

$\mathcal{L}(d|\emptyset, M_1) = \mathcal{L}(d|\varepsilon=0, \emptyset, M_2)$

A simpler model, M_1 can be recovered from a more complex model, M_2 by fixing one or more parameters.

Consistent Priors:

Priors on shared parameters of model must be consistent.

$\pi(\emptyset|M_1) = \pi(\emptyset|\varepsilon=0, M_2)$

Typically uses a seperable prior:

$\pi(\emptyset|M_1) = g(\emptyset)$ and $\pi(\varepsilon, \emptyset|M_2) = f(\varepsilon)g(\emptyset)$

Evidence for Simpler Model:

$Z_{M_1} = P(d|M_1)$ → data given model over all parameter space.

$= \int d\emptyset\, \mathcal{L}(d|\emptyset, M_1)\, \pi(\emptyset|M_1)$ → Nested models + Consistent priors

$= \int d\emptyset\, \mathcal{L}(d|\emptyset, \varepsilon=0, M_2)\, \pi(\emptyset|\varepsilon=0, M_2)$

$= P(d|\varepsilon=0, M_2)$ → Bayes Theorem

$= \frac{P(\varepsilon=0|d, M_2)}{P(\varepsilon=0|M_2)} P(d|M_2)$ → Evidence of Model 2, Z_{M_2}

$Z_{M_1} = \frac{P(\varepsilon=0|d, M_2)}{P(\varepsilon=0|M_2)} Z_{M_2}$

$B_{1,2} = \frac{Z_{M_1}}{Z_{M_2}} = \frac{P(\varepsilon=0|d, M_2)}{P(\varepsilon=0|M_2)}$

↳ Bayes Factor: $\frac{\text{Posterior PDF}(\varepsilon=0) \text{ of } M2}{\text{Prior PDF}(\varepsilon=0) \text{ of } M2}$

PDF of Prior

↳ Traditionally simple analytical function - easy to evaluate.

PDF of posterior

↳ Typically harder to evaluate

Option: MCMC to sample posterior of model, M_2: $(\varepsilon_i, \emptyset_i) \sim P(\varepsilon, \emptyset|d, M_2)$ for $i = 1, 2, \ldots, N$

Discard $\emptyset_i$ and use remaining to estimate $P(\varepsilon|d, M_2)$ - ie Kernel Desity estimation (KDE)

- Evaluate $P(\varepsilon=0|d, M_2)$

Thermodynamic Integration:

Uses Statistical Mechanics Concepts

→ annealing

→ Controlling the tempreture of the System.

$\beta^{-1} = k_B T$ → Annealing / inverse tempreture parameter, β

$0 < \beta < 1$

Annealing Likelihood $\mathcal{L}(x|d, \beta) = \mathcal{L}(d|x)^\beta$ → $\beta = 1$ (Low temp limit): Recover true likelihood → $Z(\beta=1) = Z$

$\beta = 0$ (High temp limit): Flat Likelihood (posterior ≈ prior) → $Z(\beta=0) = 1$ (prior is normalised)

Modified Posterior:

$P(x|d, \beta) = \frac{\mathcal{L}(d|x)^\beta \pi(x)}{Z(\beta)}$ → $Z(\beta) = \int_x dx\, \mathcal{L}(d|x)^\beta \pi(x)$ → β - a smooth switch between prior and posterior by turning likelihood on and off.

$$\frac{d}{d\beta}\log(Z(\beta)) = \frac{1}{Z(\beta)}\frac{dZ}{d\beta}$$

$\mathcal{L}(d|x) = \exp(\beta \log(d|x))$

$$= \frac{1}{Z(\beta)}\int_x dx\, \pi(x)\frac{d}{d\beta}\mathcal{L}(d|x)^\beta = \frac{1}{Z(\beta)}\int_x dx\, \pi(x)\,\mathcal{L}(d|x)^\beta \log \mathcal{L}(d|x)$$

$$\frac{d}{d\beta}\log(Z(\beta)) = \int_x dx\; P(x|d,\beta)\log\mathcal{L}(d|x) = E_x[\log\mathcal{L}(d|x)\,|\,\beta]$$

→ 'Expectation of $\log(\mathcal{L}(d|x))$ over all realisations of model parameters, x drawn from modified Bayesian posterior at a fixed value inverse tempreture, β'

$$\log(z) = \int_0^1 d\beta\; E_x[\log\mathcal{L}(d|x)\,|\,\beta] \qquad = [\log Z(\beta)]_0^1 = \log\left(\frac{Z(\beta=1)}{Z(\beta=0)}\right) = \log\left(\frac{Z}{1}\right)$$

↳ Requires MCMC to approximate expectations.

MCMC Implementation:

Tempreture Ladder: $(\beta_0, \beta_1, \dots, \beta_M)$

Series of increasing inverse temp: $0 \le \beta_N < 1$ for $N = 0,1,\dots,M$: $\beta_0 = 0$, $\beta_N = 1$

MCMC per Ladder Step:

For each β_N,

Run MCMC simulation to sample from $P(x|d,\beta_N)$ (annealing posterior)

N posterior Samples: $x_i^{(\beta_N)} \overset{iid}{\sim} P(x|d,\beta_N)$ for $i = 0,1,\dots,N-1$

Estimate Expectation: $E_x[\log\mathcal{L}(d|x)|\beta_N] \approx \frac{1}{N}\sum_{i=0}^{N-1}\log\mathcal{L}(d|x_i^{(\beta_N)})$

Evaluate Integral: (trapezium rule)

$$\log(z) = \frac{1}{2}\sum_{N=1}^{M}\left(E_x[\log\mathcal{L}(d|x)|\beta_N] + E_x[\log\mathcal{L}(d|x)|\beta_{N-1}]\right)\Delta\beta_N \quad \text{where } \Delta\beta_N = \beta_N - \beta_{N-1}$$

Computationally expensive: Many tempretures - β ladder steps (large M)

Many chain steps for large sample size.

Improvements:

↳ Use info from previous temp to inform starting position of next

↳ Reduce burn in period.

Parrallel-Tempered MCMC - run in parrallel and exchange info during evolution.

Nested Sampling (MC (Stochastic Sampling) algorithm)

Primarily an integration algorithm, also produces posterior samples.

→ Requires use of other stochastic sampling algorithm internally (nested).

Aim for: $Z = \int_x dx\,\mathcal{L}(d|x)\,\pi(x)$ (integral over entire parameter space, χ).

$\Xi(L)$ - 'what fraction of the prior volume has a likelihood $> L$: ie probability rv from prior gives good fit ($>L$) to data

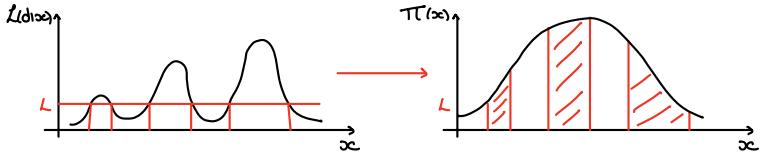
Maximum Likelihood: $\mathcal{L}_{max} = \max_{x\in\chi}\mathcal{L}(d|x)$

Prior probability/Mass: $\Xi(L) = \int_{\{x|\mathcal{L}(d|x)>L\}} dx\,\pi(x)$ Integral of Prior over region dependent on Likelihood.

Monotonically decreasing function with: $\Xi(0) = 1$, $\Xi(\mathcal{L}_{max}) = 0$

'The prior probability associated with points with a likelihood greater than a given value L'

Region of integration does not need to be simply connected ie multimodal.



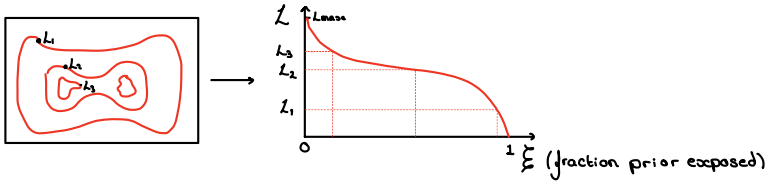
Note: both $\mathcal{L}$ and L refered to as likelihood.

Assuming x is continuous (a likelihood is not completely flat):

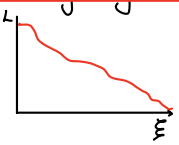
↳ $\xi(L)$ - Strictly decreasing

Thus invertible: $L(\xi)$ - Single values, well defined

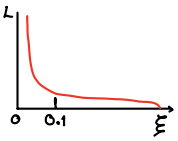
↳ $\mathcal{L}(\xi)$ - Likelihood for a given prior mass



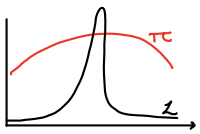
In low dimensions:



In high dimensions:



Most prior volume (parameter space) gives terrible fit to data - Likelihood - implies very well peaked likelihood relative to Prior



Example:

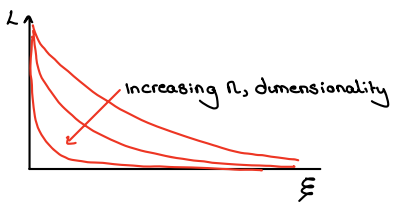
$x \in \mathbb{R}^n$ Prior: $\pi(x) = \frac{1}{V_n R^n} \begin{cases} 1 & \text{if } |x| < R \\ 0 & \end{cases}$ Likelihood: $\mathcal{L}(d|x) = \exp\left(-\frac{|x|^2}{2}\right)$

For given L: $L = \exp\left(-\frac{|x|^2}{2}\right)$ $\quad r = \sqrt{-2\ln(L)}$

Area of prior: Volume x Density $= V_n r^n \times \pi(r)$

Fraction $\xi(L) = \left(\frac{r}{R}\right)^n = \left(\frac{\sqrt{-2\ln(L)}}{R}\right)^n$

$$L(\xi) = \exp\left(-\frac{R^2}{2}\xi^{2/n}\right)$$



Considering small range of likelihood: $(L, L+dL)$

$d\xi$ - The prior mass associated with likelihood values in this range

$$d\xi = \xi(L) - \xi(L+dL) = \int_{\{x | L < \mathcal{L}(d|x) < L+dL\}} dx\, \pi(x).$$ — Prior Mass for Likelihood Values $L, L+dL$

Thus:

$Z = \int \mathcal{L}(x)\pi(x)\,dx$ — Integral of Likelihood weighted by Prior. (High dimensional

↓

$dZ = L\,d\xi$ — Discrete formulation - Likelihood, L x Prior Mass, $L+dL$

↓

$Z = \int_0^1 d\xi\, L(\xi)$ - Total Evidence - Full likelihood range. (1 Dimensional

Determining Integral Numerically:

Decreasing Sequence of M points in Prior Volume: $0 < \xi_M < \xi_{M-1} < \ldots < \xi_1 < 1$

→ With associated (threshold) likelihoods: $L_M = L(\xi_M)$

Numerical Integral (Trapezium Rule) $Z \approx \frac{1}{2}\sum_{M=1}^{M} (L_{M-1} + L_M)\,\Delta\xi_M$ where $\Delta\xi_M = \xi_{M-1} - \xi_M$ Define $\xi_0 = 1$ $L_0 = 0$.

Equivelent formulation: $Z \approx \sum_{N=1}^{M} w_N L_N$ Weights $w_N = (\xi_{N-1} - \xi_{N+1})/2$ $\xi_{M+1} = \xi_M$

Nested Sampling Algorithm

For $j=0$ to $N_{live}-1$:

$x_j \sim \pi(x)$

$L_j = \mathcal{L}(d|x_j)$

→ Initialising Parameter Space location of N_{live} live points:
by drawing independent samples from prior.
Calculate likelihood of data for each parameter values. $\mathcal{L}(d|x_j)$
N_{live} - Sufficiently large (10^3-10^4)

$Z=0$, Samples → [], $i=0$

While Condition:

$idx = \text{argmin}(L_0, \dots, L_{N-1})$

$L_i = L_{idx}$

$\xi = \exp(-i/N_{live})$

$w_i = \frac{1}{2}\left[\exp\left(-(i-1)/N_{live}\right) - \exp\left(-(i+1)/N_{live}\right)\right]$

→ For each iteration, Identify live point with lowest Likelihood.
Replace it with new sample with higher likelihood
↳ ie. draw new/replacement live point from Constrained prior
$\pi(x \mid \mathcal{L}(d|x) > L_i)$

$Z = Z + w_i L_i$

$x \sim \pi(x \mid \mathcal{L}(d|x) > L_i)$

$x_{idx} = x$

$L_{idx} = \mathcal{L}(d|x)$

→ Store discarded point - forming a list of discarded points

→ Relation (Probabilistic) between $\mathcal{L}$ and ξ

$Continue = StoppingCondition$

$i = i+1$.

→ Update Numerical Integration of evidence using trapezium weighted Sum.

Relation of L and ξ - $L(\xi)$ or $\xi(L)$ - Probabilistic

For each iteration prior mass Shrinks by $\xi_i = t\,\xi_{i-1}$

Where t random variable $(0<t<1)$ following

distribution: $P(t) = \begin{cases} N_{live}\, t^{N_{live}-1} & \text{for } 0<t<1 \\ 0 & \text{otherwise} \end{cases}$

Which has:

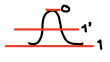
$E[\log t] = -1/N_{live}$ $\quad Var[\log t] = 1/N_{live}^2$ → This can be propogated forwards to estimate error in Evidence from Nested Sampling

$\xi_i \approx \exp(-i/N_{live})$

Justification

t - the mass fraction remains after removing lowest-likelihood sample.

For each iteration - N_{live} points sampled uniformly in remaining prior mass. (0, 1)



↳ Remaining live points - drawn from remaining mass

New threshold ξ_i - is largest among N_{live} points (largest ξ value/prior mass function).

Mathematical Derivation

N_{live}: $x_i \sim U(0,1) \rightarrow T = \max(x_i)$ $\quad P(t) = \frac{d}{dt}\int_{-\infty}^{t} dt\, P(t)$

CDF: $P(T \le t) = P(\max(x_i) \le t$ (iid)
$= P(x_1 \le t) \cdot P(x_2 \le t) \cdots P(x_{N_{live}} \le t)$
$= (P(x_i \le t))^{N_{live}} = t^{N_{live}}$

PDF: Derivative of CDF
↳ $P(t) = \frac{d}{dt} P(T<t) = \frac{d}{dt} t^{N_{live}}$

$P(t) = N_{live}\, t^{N_{live}-1}$

Stopping Criterion

NS algorithm iterates - N live points cluster to $\mathcal{L}_{peak} = \mathcal{L}_{max}$

Each Step: L_i increases and ξ_i decreases to 0

More accurate Z updates

Convergence $\Delta Z \to 0$:

Remaining Evidence Contribution

↳ $\Delta Z = L_* \xi_i$ L_* - Max $\mathcal{L}$ of live points

When $\Delta Z = L_* \xi_i < \text{tolerance}$

Stopping Condition (L_*, ξ_i, tol)

If $L_* \xi_i < tol$:

Continue = False

else:

Continue = True

Sampling from Constrained Prior:

1) Rejection Sampling from the Prior:

↳ Trivial to Sample from unconstrained prior.

Acceptance Criterion: $\mathcal{L} > \mathcal{L}'$ at sampled point.

Each Step - Prior probability of Condition decreases. (acceptance fraction).

2) MCMC (ie Metropolis Hastings)

3) Rejection Sampling from bounding distribution

4) Galilean Monte Carlo

Posterior Predictive Checks. (Sense Check)

Prior odds ratio gives prefered model from finite list of possible models.

But: Prefered Model from Subset ⟶ Correct Model

Given a model M with parameters θ and observed data d_{obs}, the Bayesian posterior can be defined: $\text{Posterior}(\theta | d_{obs}, M)$

Can define a generative model, and draw 'predicted'/unobserved data, d_{pred} from it. Such as:

Posterior Predictive Distribution

$$\theta | y_{obs}, M \sim \text{Posterior}(\theta | d_{obs}, M)$$
$$d_{pred} | \theta, M \sim \text{Likelihood}(d_{pred} | \theta, M)$$

Overall:

$$d_{pred} \sim \text{PPD}(d_{pred} | d_{obs}, \theta, M)$$

Note: Prior Predictive Distribution

↳ $\theta | M \sim \text{Prior}(\theta | M)$

$d_{pred} | \theta, M \sim \text{Likelihood}(d_{pred} | \theta, M)$

Once generated data, d_{pred}:

Compare y_{pred} with y_{obs} to check consistency

Commonly: Simple visual inspection (plots).

Flaws in PPC:

Uses y_{obs} twice: y_{pred} is generated using y_{obs}.
y_{pred} the compared with y_{obs}.

Limitations - Shouldn't use for:

⟶ Do not use to preform model comparison

Or overengineer Statistics on them

Hierachical Bayesian Models:

Gaussian Processes

An infinite-dimensional generalisation of finite-dimensional Gaussian distribution.

Recap: Univariate Gaussian Distribution

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

μ: mean

σ^2: variance (>0)

$x \sim N(\mu, \sigma^2)$

Multivariate Gaussian Distribution

$$p(\underline{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{[\underline{x}-\underline{\mu}]^T \cdot \Sigma^{-1} \cdot [\underline{x}-\underline{\mu}]}{2}\right)$$

$\underline{\mu}$ - Mean Vector ($\in \mathbb{R}^d$)

Σ - $d \times d$ Covariance matrix

(Symmetric and +ve definite)

$\underline{x} \sim N(\underline{\mu}, \Sigma)$

Note: Positive Definite

A real, Symmetric matrix Σ is positive definite if: $Z^T \cdot \Sigma \cdot Z > 0$

(Positive Semi Definite: $Z^T \cdot \Sigma \cdot Z \geq 0$) ie eigenvalues of Σ $\lambda_i > 0 \; \forall i$.

Exercise: Show that for a real, Symmetric matrix is positive definite if and only if all eigenvalues λ are strictly greater than zero, $\lambda > 0$.

For a Positive, Symmetric Matrix:

Spectral Theorem: Real, Symmetric matrix A diagonalised by Orthogonal matrix.

$A = Q \Lambda Q^T$ Q - orthogonal matrix

Λ - diagonal matrix.

$$x^T A x = x^T Q \Lambda Q^T x$$
$$= (Q^T x)^T \Lambda (Q^T x)$$
$$= y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2$$

For $x^T A x > 0$: $\sum_i^n \lambda_i y_i^2 > 0$

Gaussian Processes

defines distributions over random functions rather than finite no. random variables.

(distribution of functions).

Consider Set S, and function, f:

$f: S \to \mathbb{R}$ - maps points in set S to real numbers, $\mathbb{R}$

Sampling over functions is random (ie multiple functions randomly selected from).

↳ Gaussian Process: distribution over functions.

Advantages

- Easy interpolation in any no. dimensions / Spaces with non trivial topology.

Disadvantages

→ Computational cost

Memory Scales $\Theta(N^2)$

Mathematically:

For any Set S, a Gaussian process assigns a random value $f(x)$ to each point $x \in S$ such that for a finite $n \in \mathbb{N}$ and any collection of points $\{x_i | i = 1, \ldots n\}$ in S the vector $(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$ is distributed as a multivariate (n dim) Gaussian vector.

→ For any finite Set of inputs: $\{x_1, \ldots, x_n\}$ Corresponding outputs follow mv Gauss Dist

Distribution of functions, f from Gaussian Process, GP:

$f \sim GP(\mu, k)$ μ - Mean functions

k - Covariance functions

Set S

↳ Either finite or infinite

If S finite - GP multivariate Gaussian

Typically S continuous - $S = \mathbb{R}$

Use of Gaussian Processes:

↳ GP allows us to consider infinite dimensional smooth functions while only working with finite sets of points

Infinite Dimensional Smooth Functions ⟷ Finite Dimensional Random Vectors

Defining Gaussian Processes:

For a given set of points $(x_1, x_2, \dots, x_n)$

$$f \sim N(N(x), K(x,x')) \sim GP(N(x), K(x,x'))$$

Note: Kernel Function ≈ Covariance Function

N - mean vector → Can be arbitary.

$$N = (N(x_1), N(x_2), \dots, N(x_n))$$

K - Covariance Matrix (Kernel/Covariance function)

$$K = \begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) \\ \vdots & \ddots & \vdots \\ K(x_n,x_1) & \cdots & K(x_n,x_n) \end{pmatrix} \qquad K_{ij} = k(x_i, x_j)$$

↳ must be Symmetric positive (semi) definite function

↳ if $n \times n$ matrix with components $K_{ij} = k(x_i, x_j)$ is positive definite for all sets $n \in \mathbb{N}$ points $x_1, \dots, x_n$ in S.

↳ ie all matrices K from points in S: must be positive definite

Examples of GP:

Gaussian Process Regression.

↳ Allow easy interpolation of data in any dimension and non-trivial topologies.

Allows inclusion of prior information (as Bayesian in design).

However: Computationally expensive.

Examples of Kernel Functions:

Squared Exponential Kernel

$$K_{SE}(x,x') = \sigma_f^2 \exp\left(\frac{-(x-x')^2}{2l^2}\right)$$

$\sigma_f^2 > 0$: overall amplitude parameter

$l > 0$: lengthscale parameter

↳ Smooth - desirable property

Note: 1D definition ($S = \mathbb{R}$)

Generalised by replacing $(x-x')$ with suitable distance metric

Linear Kernel

$$K_{Linear}(x,x') = \sigma_f^2 x x'$$

$\sigma_f^2 > 0$: overall amplitude parameter

Proof of Symmetric Positive Semi-definite.

Symmetric: $K_{Linear}(x,x') = K_{Linear}(x',x)$ as $xx' = x'x$.

Semi-definite: Consider Covariance matrix $\underline{K} = \sigma_f^2 \underline{x}\,\underline{x}^T$

Eigenvalue eqn: $K e = \lambda e$

$$\sigma_f^2 \underline{x}\,\underline{x}^T e = \lambda e$$

↳ Singular eigenvector: $\underline{x}$

Singular eigenvalue: $\lambda = \sigma_f^2 \underline{x}\,\underline{x}^T$ or $\sigma_f^2 |x|^2$

↳ All the rows are linearly dependent as they a simply $\underline{x} \propto x_i$ ie vector by a scalar.

This makes it a rank = 1 matrix;

Otherwise for any eigenvector orthogonal to $\underline{x}$, v ($\underline{x}^T v = 0$)

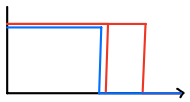
$$K \cdot v = \sigma_f^2 \underline{x}\,\underline{x}^T v = \sigma_f^2 \underline{x} \cdot 0 = 0.$$ Thus eigenvalue 0.

Brownian Kernel:

$$K_{Brownian}(x,x') = \sigma_f^2 \min(x,x')$$

Proof of Positive definite Kernel function:

$K_{Brownian}$ rewritten: $\boxed{K_{Brownian}(x,x') = \sigma_j^2 \int_0^\infty dt\, H(t;x)\, H'(t;x')}$ where $H(t;x) = \begin{cases} 1 & t \le x \\ 0 & t > x \end{cases}$



Aim to show non-negative for any set of points $(x_1, x_2, \dots x_N)$

for any vector $z \in \mathbb{R}^N$

$$z^T \cdot K \cdot z = \sum z_i z_j K_{ij}$$

$$= \sum_{ij} \sigma_j^2 z_i z_j \int_0^\infty dt\, H(t;x_i)\, H(t;x_j)$$

$$= \sigma_j^2 \int_0^\infty dt \underbrace{\sum_{i,j} z_i z_j H(t,x_i) H(t,x_j)}_{\left(\sum_i z_i H(t,x_i)\right)\left(\sum_j z_j H(t,x_j)\right) = \left(\sum z_i H(t,x_i)\right)^2}$$

$$= \sigma_j^2 \left(\sum z_i H(t,x_i) \right)^2 \geq 0.$$

Covariance Functions:

Given an example 'Seed Covariance function', it can be adjusted accordingly:

Summation:

For any two Covariance functions: $K_1(x,x')$ and $K_2(x,x')$

Their Sum: $\boxed{K(x,x') = K_1(x,x') + K_2(x,x')}$ is also a Covariance function

Product:

For any two Covariance functions $K_1(x,x')$ and $K_2(x,x')$

Their product: $\boxed{K(x,x') = K_1(x,x')\, K_2(x,x')}$ is also a covariance function.

Warping:

For any Covariance function $K(u,u')$ on the Space S' and a warping function $U: S \to S'$: the function $\boxed{\mathcal{K}(x,x') = K(U(x), U(x'))}$ is a new covariance function on the Space S. (U pulls back Kernel S' to S). Where $x, x' \in S$, $U, U' \in S'$

Renormalisation:

For any Covariance function $K(x,x')$ and function $\alpha: S \to \mathbb{R}$ the function $\boxed{\alpha(x)\,\alpha(x')\,K(x,x')}$ is a new Covariance function

Proof of warping function:

Show $\mathcal{K}(x,x')$ is positive definite: $z^T \cdot K \cdot z = \sum_{ij} z_i z_j \mathcal{K}_{ij}$

$$= \sum_{ij} z_i z_j \mathcal{K}(x_i, x_j) \quad \text{for any choice } x_i$$

$$= \sum_{ij} z_i z_j K(U(x_i), U(x_j))$$

$$= \sum_{ij} z_i z_j K(u_i, u_j) \quad \text{Where } u_i = U(x_i)$$

$$\geq 0 \text{ as } K \text{ is positive semi definite.}$$

Example with Periodic Kernel:

Let $S' = \mathbb{R}^2$ Using 2D Square Exp Kernel $R\left(\begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} x' \\ y' \end{pmatrix}\right) = \exp\left(\frac{-(x-x')^2 - (y-y')^2}{2l^2}\right)$

Trying to enduce/Pull back Gaussian Process onto Unit Circle/ New Space:

$S = S_1$ (circle)

$U(\phi) = \begin{pmatrix} \cos(\phi) \\ \sin(\phi) \end{pmatrix}$ Define a new Kernel K on S_1 by

$$K(\phi, \phi') = k(U(\phi), U(\phi'))$$

$$= \exp\left(-\frac{1}{2l^2} \left[(\cos(\phi) - \cos(\phi'))^2 + (\sin(\phi) - \sin(\phi'))^2 \right] \right)$$

$$\left[(\cos(x) - \cos(x'))^2 + (\sin(x) - \sin(x'))^2 = 4\sin^2\left(\frac{x - x'}{2}\right) \right]$$

$$= \exp\left(-\frac{2}{l^2} \sin^2\left(\frac{\phi - \phi'}{2}\right) \right)$$ Periodic Kernel

↳ proved to be Semi definite assuming Squared exponential

Focusing on the Real Line: $S = \mathbb{R}^n$

Stationary Kernel

Only depends on the difference of the inputs:

$\boxed{K(x, x') = K(\tau)}$ where $\tau = x - x'$

A Stationary Kernel is invariant under translations

$\boxed{k(x, x') = k(x + \Delta, x' + \Delta)}$ for all $\Delta \in \mathbb{R}^n$.

Squared Exponential Kernel:

↳ Both Stationary and isotropic.

Isotropic Kernel

a Stationary distribution that only depends on Euclidean Distance

ie $k(\tau) = k(r)$ $r = |\tau|$

Invariant under both translations and rotations

Properties of Stationary Kernels:

i) By Symmetry property: $K(x, x') = K(x', x)$

↳ $K(-\tau) = K(\tau)$

ii) Positivity: $K(\tau = 0) = K(x, x) \geq 0$

iii) boundedness: $K(\tau) \leq K(\tau = 0)$ for all $\tau \in \mathbb{R}^d$

$K(\tau) = \begin{pmatrix} K(0) & K(\tau) \\ K(\tau) & K(0) \end{pmatrix}$ → Eigenvalues to be positive:

must be: $K(\tau) < K(0)$

Proving Stationary Kernels are positive (Semi) definite:

Lemma: A function $k: \mathbb{R}^n \to \mathbb{R}$, denoted $K(\tau)$ can be used to define a Semi-definite Stationary Kernel: $k: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as $K(x, x') = k(x - x')$ if the fourier transform $\tilde{K}(f) \geq 0$ for all $f \in \mathbb{R}^n$

Fourier transform:

$$\tilde{K}(f) = \int K(\tau)\, e^{(2\pi i f \cdot \tau)}$$

Proof: Suppose the fourier transform always exists:

$$K(\tau) = \int_{\mathbb{R}^n} df\, \tilde{k}(f) \exp(-2\pi i f \cdot \tau)$$

Now Considering:

$$z^T \cdot K \cdot z = \sum z_i z_j K(x_i, x_j)$$ ↓ Stationary Kernel

$$= \sum z_i z_j K(x_i - x_j)$$

$$= \sum_{ij} z_i z_j \int df\, \tilde{k}(f) \exp(-2\pi i f \cdot (x_i - x_j))$$ ↓ As $\tau = x_i - x_j$

$$= \int df \sum_{ij} z_i z_j \tilde{k}(f) \exp(-2\pi i f x_i) \exp(2\pi i f x_j)$$

Complex conjugate

↳ **Generalised Result of Bochner's Theorem**

A continuous function $k(\tau)$ is positive Semi definite if and only if it is the Fourier transform of a finite non-negative Borel measure N on $\mathbb{R}^d$.

ie. $K(\tau) = \int_{\mathbb{R}^n} dN(f) \exp(-2\pi i f \cdot x)$

$$= \int df\ \tilde{K}(f) \underbrace{\left| \sum z_i \exp(-2\pi i f \cdot x_i) \right|^2}_{\geq 0} \geq 0 \quad \text{if } \tilde{K}(f) \geq 0$$

Thus if $\tilde{K}(f) \geq 0$ for all f: The Stationary Kernel is positive Semi definite.

The Squared Exponential Kernel:

We first Show it is a Stationary Kernel, $K_{SE} = \sigma_f^2 \exp\left(\frac{-(x-x')^2}{2l^2}\right) = \sigma_f^2 \exp\left(\frac{-\tau^2}{2l^2}\right)$

ie $K(x,x') = K(x+\Delta, x'+\Delta)$

Taking its fourier transform: $\tilde{K}(f) = \int_{-\infty}^{\infty} \sigma_f^2 \exp\left(\frac{-\tau^2}{2l^2}\right) \exp(2\pi i f \tau)\, d\tau$

$$= \sigma_f^2 \int_{-\infty}^{\infty} \exp\left(-\frac{\tau^2}{2l^2} + 2\pi i f \tau\right) d\tau$$

Completing the Square: $-\frac{\tau^2}{2l^2} + 2\pi i f \tau = -\frac{1}{2l^2}\left[\tau^2 + 4\pi i f l^2 \tau\right]$

$$= -\frac{1}{2l^2}\left(\left[\tau + 2\pi i f l^2\right]^2 - 4\pi^2 f^2 l^4\right)$$

$$= -\frac{1}{2l^2}\left[\tau + 2\pi i f l^2\right]^2 - 2\pi f^2 l^2$$

$$= \sigma_f^2 \exp\left(2\pi^2 f^2 l^2\right) \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2l^2}\left[\tau + 2\pi i f l^2\right]^2\right\} d\tau$$

$$= \sigma_f^2 \exp\left(2\pi^2 f^2 l^2\right) \underbrace{\int_{-\infty}^{\infty} \exp\left\{-\frac{u^2}{2l^2}\right\} du}_{l\sqrt{2\pi}\ :\ \text{Gaussian Integral}} = \sigma_f^2 \exp\left(2\pi^2 f^2 l^2\right) \times l\sqrt{2\pi}$$

$$\tilde{K}(f) = \sigma_f^2\, l\sqrt{2\pi} \exp\left(-2\pi^2 f^2 l^2\right) \geq 0 \text{ for all } f$$

Note: Gaussian Integral

$$\int_{-\infty}^{\infty} \exp\left(\frac{-u^2}{2\sigma^2}\right) du = \sigma\sqrt{2\pi}$$

Gaussian Process Regression:

(Interpolation + Extrapolation)

Lemma: Let $x \sim N(\mu, \Sigma)$. Split the vector $x \in \mathbb{R}^n$ into $x^T = (x_1^T, x_2^T)$

where $x_1 \in \mathbb{R}^{n_1}$ and $x_2 \in \mathbb{R}^{n_2}$ and $n_1 + n_2 = n$.

Similarly Spit $\mu^T = (\mu_1^T, \mu_2^T)$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ where $\Sigma_{11} - n_1 \times n_1$, $\Sigma_{12} - n_1 \times n_2$, $\Sigma_{22} - n_2 \times n_2$

The distribution of x_1 conditioned on a fixed value of x_2 is:

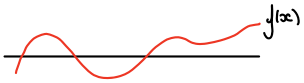
$x_1 | x_2 \sim N(\mu', \Sigma')$ where $\mu' = \mu_1 + \Sigma_{12} \cdot \Sigma_{22}^{-1} \cdot (x_2 - \mu_2)$

$\Sigma' = \Sigma_{11} - \Sigma_{12} \cdot \Sigma_{22}^{-1} \cdot \Sigma_{12}^T$

} Distribution of Subset$_1$ With respect to fixed Subset$_2$

In words - If you take a Slice through an n-dimensional Multivariate Gaussian along any m-dimensional hyperplane ($m < n$) then the pdf along this Slice has the Shape of another Gaussian: $x_1 | x_2$

Tackling Regression:



Suppose we have measured the value of the function at n points, $\underline{x} = (x_1, x_2, \dots, x_n)$ and the measured values are $f = (f(x_1), \dots, f(x_n))$ and we wish to predict the value $f(x_*)$ at some new location x_*.

Prior on Space of Functions: Use Zero-mean Gaussian Process (assume you have scaled to zero-mean)

$f \sim GP(0, k)$ → Any General Kernel.

Def of Gaussian Process:

apply to $(x_*, x_1, \dots, x_n)$

Prior means:

$$(f(x_*), f_1, \dots f_n) \sim N(\underline{0}, K')$$

Where $K' = \begin{pmatrix} K_{**} & K_*^T \\ K_* & K \end{pmatrix}$

Where $\underline{k}_{**} = k(x_*, x_*)$ → Single value, $\mathbb{R}$

$\underline{k}_* = k(x_*, x_i)$ → Vector, $\mathbb{R}^n$ $i = 1, \dots, N$ $= (k(x_*, x_1), k(x_*, x_2), \dots, k(x_*, x_n))$

$K = K(x_i, x_j)$ $i, j = 1, \dots, N$ → Matrix, $\mathbb{R}^{n \times n}$ $= \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$

$$f(x_*), f \sim N\left(0, \begin{pmatrix} K_{**} & K_*^T \\ K_* & K \end{pmatrix}\right)$$

Determining the Conditional distribution:

$$f(x_*) \mid f = N(E(f(x_*)), \text{Var}(f(x_*)))$$

$$E[f(x_*)] = k_*^T \cdot K^{-1} \cdot f$$
$$\text{Var}[f(x_*)] = k_{**} - k_*^T \cdot K^{-1} \cdot k_*$$

View point: 'Gaussian Processes are a type of Conjugate Prior

Prior: $f \sim GP(0, k)$ → This Conjugate Prior as Posterior distributed in the same way: Gaussian Process.

Posterior: $f(x_*) \mid f \sim N(E[f(x_*)], \text{Var}[f(x_*)])$

Example Conjugate prior with update of parameters

$N \to N'$: $E[f(x_*)] = k_*^T \cdot K^{-1} \cdot f$

$k \to k'$: $\text{Var}[f(x_*)] = k_{**} - k_*^T \cdot K^{-1} \cdot k_*$

Considering Measurement Values, $f(x_i)$, with associated Gaussian errors, σ_i

Adding a diagonal term to Covariance Matrix.

$\sigma^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)^T$

$$E[f(x_*)] = k_*^T \cdot (K + \text{diag}(\sigma^2))^{-1} \cdot f$$
$$\text{Var}[f(x_*)] = K_{**} - K_*^T \cdot (K + \text{diag}(\sigma^2))^{-1} \cdot K_*$$

→ Additionally done to help Numerical Stability.

Evidence of Gaussian Process.

$$\log(Z) \equiv \log P(y) = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi)$$

↳ Can use this to help inform Covariance function.

Gaussian Process Review: ('mean Square': Continuity and Differentiability).

Properties of Gaussian Process are determined by k:

'terms like continuity etc become hard to define traditionally as we are discussing finite sampeled points'

Restrict $S \in \mathbb{R}^n$, Let $x_1, x_2, \ldots$ approach/converge x_* if $|x_i - x_*| \to 0$ as $i \to \infty$

Mean Square Continuous

f is mean Square continuous if $E[|f(x_*) - f(x_i)|^2] \longrightarrow 0$ as $i \to \infty$

If this holds for all $x_* \in \mathbb{R}^n$ then we say f is MS continuous everywhere

Relating this condition to Kernel function:

f is Mean Square continuous at x_* iff Kernel function $k(x, x')$ is 'Continuous' ↳ Usual Sense.

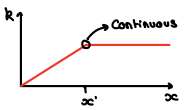
at $x = x' = x_*$ (and $\mu(x)$ is continuous at $x = x_*$).

For a Continuous Kernel: $k(\tau) = k(x, x')$ where $\tau = x - x'$

Simply requires ensuring $k(0)$ is continuous (translation invariance). → Can show it once.

$k_{SE} = \sigma_f^2 \exp\left(-\frac{\tau^2}{2l^2}\right)$ Continuous $\tau = 0$

$k_{Brownian}(x, x') = \sigma_f^2 \min(x, x')$ also continuous.



Mean Square Differentiability

$$\frac{\partial f}{\partial x_i} = \text{Limit in MS}\left(\frac{f(x + \varepsilon e_i) - f(x)}{\varepsilon}\right)$$

If $f \sim GP$ then if it exists $\frac{\partial f}{\partial x_i} \sim GP$ With $K(x, x') = \frac{\partial^2 K(x, x')}{\partial x_i \partial x_i'}$

The level of mean Square differentiable of f is determined by level of differentiability of kernel k (with factor of k)

$k_{Brownian}$ - MS continuous but not differentiable.

Often k_{SE} - Over Smooths:

Alternative: Matern Kernel Function

$$k(x, x') = k(t) = \frac{\sigma_f^2 2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\tau}{l}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{\tau}{l}\right)$$ → modified Bessel Function

Considered Generalisation k_{SE}

$\nu \to \infty$: Matern → SE

ν controls 'level of Smoothness'.

Derivatives at $\tau = 0$; This kernel is α times MS differentiable if $\nu > \alpha$. as kernel is $\text{int}(\alpha)$ time differentiable

Computational Cost of GPR:

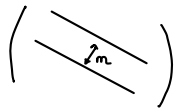
K^{-1} has dimensions $n \times n$: $\Theta(n^2)$ memory

$\Theta(n^3)$ Computation

Compact Support kernel:

kernels such as k_{SE} require dense matrix ie all positions have values (but many very small).

Wendland Polynomials:

Leaves Band Diagonal K  memory $\Theta(n \times m)$

Hierarchical Bayesian Models (and Probabilistic Graphical Models.)

Parameters split across several 'levels', reflected in a particular factorisation of the prior.

Example Problem 1:

Estimating the age of a single star: τ using observations: t,

Regular Bayesian Inference:

→ Gaussian Prior: $\frac{1}{\sqrt{2\pi\Delta^2}} \exp\left(-\frac{1}{2}\frac{(N-\tau)^2}{\Delta^2}\right)$

Gaussian likelihood. $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(t-\tau)^2}{\Delta^2}\right)$

Alternatively $\tau \mid N, \Delta \sim N(N, \Delta^2)$

$t \mid \tau, \sigma \sim N(\tau, \sigma^2)$

Prob Graphical Model

- A way of illustrating joint distributing on every random variables. (Directed AcyclicGraph - DAG)

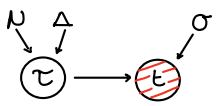
Fixed values: N, Δ, σ

↳ No circles

Random variables: t, τ

↳ Shaded: observed variables

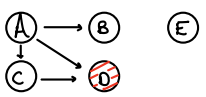
Empty: parameters/latent variables



Arrow: Denote order of conditional probabilities.

Ie generating fake data (what order?)

Example Problem 2:



$P(A,B,C,D,E) = P(A)\,P(B|A)\,P(C|A)\,P(D|A,C)\,P(E)$

Can also write: (for observed D)

$P(A,B,C,D,E) = P(A,B,C,D,E|D)\,P(D)$

$P(A,B,C,D,E|D) = \frac{1}{P(D)}\,P(D|A,C)\,P(A)\,P(B|A)\,P(C|A)\,P(E)$

Posterior. | Evidence | likelihood | Prior

Bayesian Posterior:

$$P(\tau|t) = \frac{1}{P(t)} P(\tau)\,P(t|\tau)$$

$$= \frac{1}{Z} \pi(\tau)\,P(t|\tau)$$

Where Evidence $Z(\sigma, N, \Delta) = \int d\tau\, \mathcal{L}(t|\tau,\sigma)\,\pi(\tau|N,\Delta)$

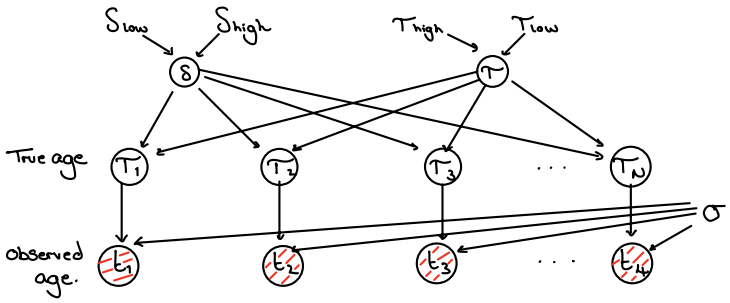
Example Hierarchical Model:

Assume all stars in a cluster formed around the same time.

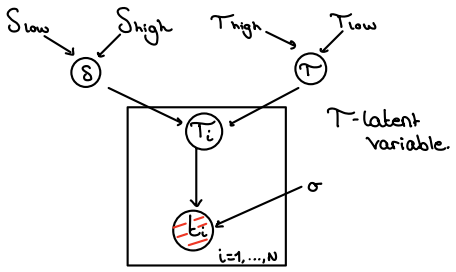
Hyper Prior: $\begin{cases} \tau \mid \tau_{low}, \tau_{high} \sim \text{log uniform}(\tau_{low}, \tau_{high}) \\ \delta \mid \delta_{low}, \delta_{high} \sim \text{log uniform}(\delta_{low}, \delta_{high}) \end{cases}$

Prior $\tau_i \mid \tau, \Delta \sim N(\tau, \delta^2)$ δ- Small scatter in ages within cluster

Likelihood: $t_i \mid \tau_i, \sigma \sim N(\tau_i, \sigma^2)$ independently for $i = 1, \ldots, N$



or alternatively



Bayes Theorem:

→ Neglecting fixed constants

Eq 1: $$P(\tau, \delta, \{\tau_i\}, \{t_i\}) = P(\tau)\,P(\delta)\,P(\{\tau_i\}|\delta,\tau)\,P(\{t_i\}|\{\tau_i\})$$

$$= \underbrace{\pi(\tau)\,\pi(\delta)}_{\text{Hyper Priors}} \underbrace{\prod_i^{N_s} \pi(\tau_i|\delta,\tau)}_{\text{Prior}} \underbrace{\prod_i^{N_s} \mathcal{L}(t_i,\tau_i)}_{\text{Likelihood.}}$$

Eq 2:

$$P(\tau, \delta, \{\tau_i\}, \{t_i\}) = P(\{t_i\})\,P(\tau, \delta, \{\tau_i\}|\{t_i\})$$

Equating with posterior:

$$\underbrace{P(\tau, \delta, \{\tau_i\}|\{t_i\})}_{\text{Posterior}} = \underbrace{\frac{1}{P(\{t_i\})}}_{\text{Evidence}} \underbrace{\pi(\tau)\,\pi(\delta)}_{\text{Hyper Priors}} \underbrace{\prod_i^{N_s} \pi(\tau_i|\delta,\tau)}_{\text{Prior}} \underbrace{\prod_i^{N_s} \mathcal{L}(t_i,\tau_i)}_{\text{Likelihood.}}$$

Truely we are intrested in τ, δ (not individual τ_i)

Marginalise over latent variables - τ_i

$$P(\tau,\delta|\{t_i\}) = \frac{1}{P(\{t_i\})}\,\pi(\tau)\,\pi(\delta)\,\boxed{\prod_{i=1}^{N_s} \int d\tau'\,\pi(\tau'|\delta,\tau)\cdot\mathcal{L}(t_i|\tau')}$$

→ Product of evidence of Single models.