- To get frequentist properties of estimator have to draw from a distribution, can test diff priors vs diff distrib.

- Using priors : encode background information

⇒ also want to test sensitivity of inferences to model

POSTERIOR EXPECTATION $\mathbb{E}[f(\underline{\theta}) | D)] = \int f(\underline{\theta}) P(\underline{\theta} | D) d\underline{\theta}$

This is what we want to get, but computationally difficult

⇒ here Bayesian computation methods

"map out" / sample posterior density to compute expectation

Multi-parameter Bayesian inference

Gaussian example : Gelman BDA sec 3.2-3.3

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

so $P(\underline{y} | \mu, \sigma^2) = \prod_{i=1}^{n} N(y_i | \mu, \sigma^2) = (\sigma\sqrt{2\pi})^{-n} \exp\left\{ -\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2] \right\}$

(s^2: sample variance of y; $\bar{y}$: sample mean of y)

- Prior : $P(\mu, \sigma^2) = P(\mu | \sigma^2) P(\sigma^2)$

$\mu | \sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$ (how much information have about σ^2)

and $\sigma^2 \sim \text{Inv}\chi^2(\nu_0, \sigma_0^2)$ inverse χ^2 distribution

$P(\mu, \sigma^2 | \underline{y}) \propto P(\underline{y} | \mu, \sigma^2) P(\mu | \sigma^2) P(\sigma^2)$

$= N(\mu | \mu_n, \sigma_n^2/\kappa_n) \cdot \text{Inv}\chi^2(\sigma^2 | \nu_n, \sigma_n^2)$ combine information of y and prior

where $\kappa_n = \kappa_0 + n$, $\nu_n = \nu_0 + n$, $\mu_n = \frac{\kappa_0}{\kappa_0 + n} + \frac{n\bar{y}}{\kappa_0 + n}$, $\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu)^2$

$n \to \infty$ $\quad \mu_n \to \bar{y} \quad \sigma_n^2 \to s^2 \quad \kappa_n \to n \quad \nu_n \to n$

so $P(\mu, \sigma^2 | y) = N(\mu | \bar{y}, \sigma_n^2) \cdot \text{Inv}\chi^2(\sigma^2 | n, s^2)$

↑ similar to inverse Γ

now integrate out variables:

$$P(\sigma^2 | y) = \int P(\mu | \sigma^2, y) P(\sigma^2 | y) \, d\mu = \text{Inv}\chi^2(\sigma^2 | \nu_n, \sigma_n^2)$$

$$P(\mu | y) = \int P(\mu, \sigma^2 | y) \, d\sigma^2 = \ldots \propto \left[1 + \frac{\kappa_n (\mu - \mu_n)^2}{\nu_n \sigma_n^2} \right]^{-(\nu_n + 1)/2}$$

$$= t_{\nu_n}(\mu | \mu_n, \sigma_n^2 / \kappa_n)$$ t distribution!

and as $n \to \infty \rightarrow t_n(\mu | \bar{y}, s^2/n) \to N(\mu | \bar{y}, s^2/n)$

t distribution goes to gaussian

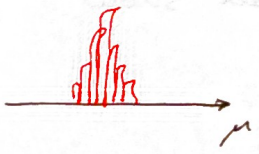
- What do we do if we can't derive this?

⇒ <u>Monte Carlo direct sampling</u>

use factorising, ie that we know posterior is normal × something

draw from both factors of posterior

and make histogram of posterior samples



$\theta_i \sim P(\theta | D) \qquad \hat{I} = \frac{1}{m} \sum_{i=1}^{m} f(\theta_i) \rightarrow$ expectation for large m

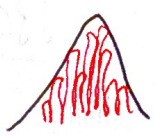
- f could be mean, variance, posterior prob over interval

<u>NB</u> how posterior expectation is <u>not</u> averaging over data but over parameters

- for above: $\sigma_i^2 \sim P(\sigma^2 | y)$, $\mu_i | \sigma_i^2 \sim P(\mu | \sigma^2, y)$ } look at joint distribution of samples

KDE kernel density estimate <u>smooths</u> density estimate

put a gaussian + bandwidth on each point ⇒ add them all up



- What if you can't sample posterior? MCMC, Nested, etc. for long-run generation

Importance sampling

draw from a simpler distribution

- have probability density $P(\theta)$ ← intractable — can write down but can't easily draw sample

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

- make $Q(\theta)$ instrumental distrib. / importance function

easy to draw from $\theta_i \overset{iid}{\sim} Q(\theta)$

important: $Q(\theta) > 0$ wherever $P(\theta) > 0$

$$I = \mathbb{E}_P[f(\theta)] = \int f(\theta) \frac{P(\theta)}{Q(\theta)} Q(\theta) d\theta \simeq \frac{1}{m} \sum_{i=1}^{m} \frac{P(\theta_i)}{Q(\theta_i)} f(\theta_i)$$

$$\hat{I} \simeq \frac{1}{m} \sum_{i=1}^{m} \underbrace{W(\theta_i)}_{\text{IMPORTANCE WIDTH } \frac{P(\theta_i)}{Q(\theta_i)}} f(\theta_i)$$

$\mathbb{E}_Q$ expectation over $Q(\theta)$

$\mathbb{E}_P$ expectation over $P(\theta)$

- Can show $\mathbb{E}_Q[\hat{I}] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_Q[f(\theta_i) W(\theta_i)]$

$$= \frac{1}{m} \sum_{i=1}^{m} \underbrace{\mathbb{E}_P[f(\theta_i)]}_{I} = I$$

(1) Self Normalised

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

$P(D)$ is called evidence / partition function $Z_P = \int \tilde{P}(\theta|D) d\theta$

- often only have $\tilde{P}(\theta|D) = \underbrace{P(D|\theta)P(\theta)}_{\equiv \tilde{P}(\theta) \text{ below}}$ don't know $P(D)$!

- Similarly to above:

$$I = \mathbb{E}[f] = \int f(\theta) \underbrace{P(\theta|D)}_{\frac{\tilde{P}(\theta|D)}{\int \tilde{P}(\theta|D) d\theta}} d\theta = \int f(\theta) \frac{\tilde{P}(\theta|D)}{\int \tilde{P}(\theta|D) d\theta} d\theta$$

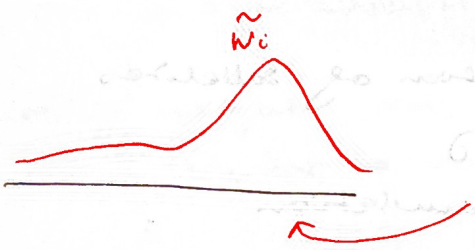
$$= \frac{\int f(\theta) \tilde{P}(\theta)/Q(\theta) \, Q(\theta) \, d\theta}{\int \frac{\tilde{P}(\theta)}{Q(\theta)} Q(\theta) \, d\theta}$$

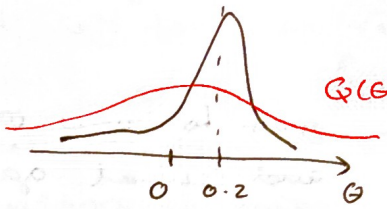
$$\Rightarrow \hat{I} = \frac{\sum_{i=1}^{m} f(G_i)\,\tilde{W}(G_i)}{\sum_{i=1}^{m} \tilde{W}(G_i)} \qquad \text{where } \tilde{W}_i = \frac{\tilde{P}(G_i)}{Q(G_i)} \qquad G_i \overset{iid}{\sim} Q(G), \; i = 1, \dots, m$$

$$= \sum_{i=1}^{m} f(G_i) W(G_i) \qquad W_i = \frac{\tilde{W}_i}{\sum_{i=1}^{m} \tilde{W}_i} = \frac{\tilde{P}(G_i)/Q(G_i)}{\sum_{i=1}^{m} \tilde{P}(G_i)/Q(G_i)}$$

SELF-NORMALISED IMPORTANCE WEIGHT

- Example: two gaussians





from Q too many samples of $Q > P$, too little samples $Q < P$

overweight samples where $Q < P$

- Example: in parallax example can't compute physical prior can't normalise posterior

with these $\sum x_i W_i$ get estimate of mean

(mode was easier analytically)

NB error doesn't go down as $\sqrt{N}$ bc samples are weighted. (from G, not P)

- choosing G: theoretically want $Q = \frac{f(G)\,P(G)}{\int f(G)\,P(G)\,dG}$ but usually not possible

- ν const W_i otherwise get too large variance
- sample size is also diff bc of weighting, interested by distribution of weighting

→ find thick-tailed distribution infinities make variances large

Monte Carlo Integration

m samples from posterior $\theta_i \sim P(\theta|D)$

$$\hat{I} = \frac{1}{m} \sum f(\theta_i) \longrightarrow I \quad \text{by LLN}$$

error: $\mathrm{Var}[\hat{I}] = \frac{1}{m}\mathrm{Var}[f(\theta)] \propto \frac{1}{m}\widehat{\mathrm{Var}}[\sum f(\theta_i)]$

Milky Way Mass

dwarf satellite galaxies
look at their mass, momentum, position

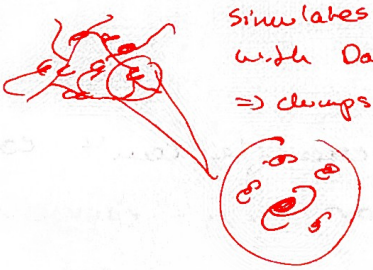
satellites

MW central galaxy

want to know correlation of satellites and central galaxy

⇒ get prior from a simulation

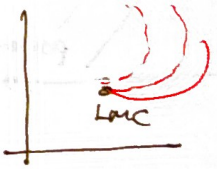
illustris



simulates a box of the universe with Dark Matter + normal matter
⇒ clumps into galaxies

look at correlation for all these simulated

- measure velocity, trajectory of surrounding dwarf galaxies
  ⇒ past trajectory depends heavily on MW mass



- setup: $\underline{x}$ true values of v, r, j ← momentum (velocity, position) } $\underline{\theta} = (\underline{x}, M_{vir})$
  M_{vir} mass of central galaxy

our prior can't be evaluated! only simulated

- measurement error likelihood: $\mathcal{L}(\underline{x}|\underline{d})$ some normal errors

  $\underline{d}$ are measurements of $\underline{x}$

- $P(\underline{x}, M_{vir}|\underline{d}) \propto P(\underline{d}|\underline{x}) \cdot P(\underline{x}, M_{vir})$
  ↑ observations, ↑ prior

have samples from prior as points in simulation

$$\int f(\theta) P(\underline{x}, M_{vir}|\underline{d}) d\theta \propto \frac{\sum f(\theta_j) P(\underline{d}|\underline{x}_j)}{\sum P(\underline{d}|\underline{x}_j)}$$

- In this case ζ only depends on M_{vir}

$$\int \zeta(M_{vir}) P(M_{vir} | \underline{d}) \, dM_{vir} \approx \sum \zeta(M_{vir,j}) w_j$$

$$w_i = P(\underline{d} | \underline{x}_i) \; / \; \sum P(\underline{d} | \underline{x}_j)$$ proportional to likelihood of each ~~parameter~~ sample

here no choice for Φ, given by simulation

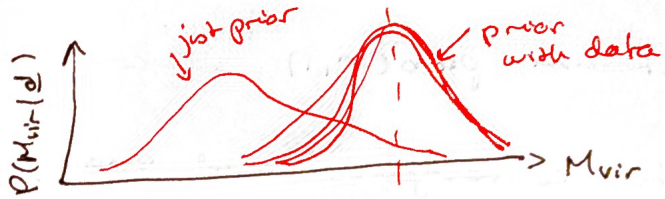
⇒ never have to evaluate it because it cancells out!

- importance weight: only a few values contribute!
  systems similar to our galaxy contribute, others not much

⇒ can use these to calculate a <u>weighted KDE</u>

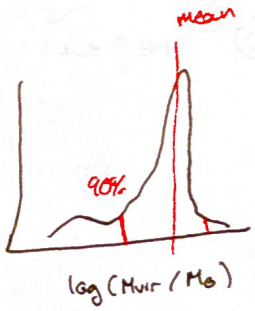
$$\text{whole}(\Theta) = \sum w_s \cdot N(\Theta | \Theta_s, bw^2)$$ ← some rule of thumb → bandwidth

make graphs like:



get MW:

$M_{vir} = 1.7^{+1.33}_{-0.52} \cdot 10^{12}$



- can calculate highest posterior density intervals

<u>NB</u> in paper don't use log so by Jensen's inequality get diff value for $\log M_{vir}$ than from this

- using multiple satellites get a better estimate

Markov Chain Monte Carlo

need to use this for higher number of parameters

- need to generate samples from a posterior without Φ
- generate a chain of RVs that in a limit are draws from poste

next value depends on current value! correlated

Metropolis

① choose μ_0

② propose $\mu_{prop} \sim N(\mu_i, \Sigma^2)$ ← choose cleverly

③ evaluate $$r = \frac{P(\mu_{prop} | y)}{P(\mu_i | y)}$$

④ $r \geq 1$ μ_{prop} better $\mu_{i+1} = \mu_{prop}$

else accept with prob$(r, 1)$ so could accept even if not better

⑤ repeat until convergence and enough data