A. VINCENT
7/3/19
QUEEN'S U

# So you want to be a Bayesian

Statistics is what we use to test a hypothesis when we have incomplete information.

References: Trotta 1701.01467 (mainly)
Gelman et al. Bayesian Data Analysis 3rd ed.
Argüelles & Colin Stats for Little Penguins

## 1) Probability

The definition of probability that we are taught in kindergarten is the following:

> $P(A)$ is the fraction of times the outcome of an experiment, performed $N \gg 1$ times, is A.

This is the <u>frequentist</u> interpretation.

The main tool is the likelihood $\mathcal{L}(\theta) \equiv P(d|\theta)$ (data $d$, parameter $\theta$)

When evaluating confidence intervals, p-values, etc, the frequentist asks: how (un)likely is this data set assuming my model $\mathcal{M}(\theta)$ is true?

The important caveat is that $\mathcal{L}$ is <u>not</u> a P.D.F. in $\theta$ $\Rightarrow$ p-value $\neq$ probability ~~of~~ the null hypothesis, but we would really like it to be. This leads to the Bayesian's question:

- Given my current state of knowledge, how confident can I be that $\mathcal{H}(\theta)$ is true?

✓

Some definitions: $A$ : some value
$\bar{A}$ : not $A$

$P(A)$ : prob. of $A$

$\Rightarrow P(A) + P(\bar{A}) = 1$

$P(A,B)$ : Joint probability (P of A and B)

$P(A|B)$ : Conditional probability (P of A given B)

Note $P(A,B) \equiv P(A|B)\,P(B)$

(Since $P(A,B) = P(B,A)$

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

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

↑ Bayes' theorem

In the language of inference:

$$\underbrace{P(\theta|d)}_{\text{posterior probability}} = \frac{P(d|\theta)\,P(\theta)}{P(d)}$$

$P(d|\theta) \rightarrow \mathcal{L}(\theta)$

$P(\theta) \rightarrow$ prior probability

$P(d) \rightarrow = Z$, evidence

$\mathcal{L}$ : the same as in the frequentist def.

Noting that $P(A) = P(A,B_1) + P(A,B_2) + P(A,B_3) + \cdots$

$$= \sum_i P(A|B_i)\,P(B_i)$$

or in the continuous case $P(A) = \int dB\, P(A|B)\,P(B)$

model ↑

We can define the evidence $P(d)$ (really $P(d|\mathcal{M})$)

$$P(d) = \int d\theta\, P(d|\theta) P(\theta) \equiv Z(\mathcal{M})$$

This is conceptually straightforward, but usually very difficult to compute. We will (maybe) get back to that.

$P(\theta)$: Prior (probability): we interpret this as our knowledge of the theory before looking at the data.

Why bother with all this?

- Frequentist methods often rely on asymptotic properties of estimators, so you need to be very careful (e.g. Wilks' theorem).

- Bayes stats deal trivially with nuisance parameters

  Say $\eta$ is a background parameter, but we only care about $\theta$:

  $$P(\theta) = \int P(\theta, \eta)\, d\eta$$

  In the MCMC approach this is just a sum.

- Prior info can be included in a meaningful way

In practice, the form of Bayes' thm makes it very easy to obtain $P(\theta|d)$.

## 2. The Bayesian method

1) Choose a model and write the parameters as a vector $\vec{\theta}$

eg $\Lambda CDM$ : $\theta = \{ H_0, \tau, \Omega_m, \Omega_b, A_s, n_s \}$

2) Specify priors. For example if I am using CMB data, $P(H_0)$ could be a gaussian reflecting HST SNe measurements.

Priors should include relevant ext. information

3) Construct $\mathcal{L}(\theta)$. This reflects the way in which data are acquired. For example Gaussian noise → normal distribution, counting → Poisson, etc.

Nuisance parameters such as background rates should be included.

4) Obtain the posterior. For simple distributions this is analytical. Usually methods such as MCMC are required.

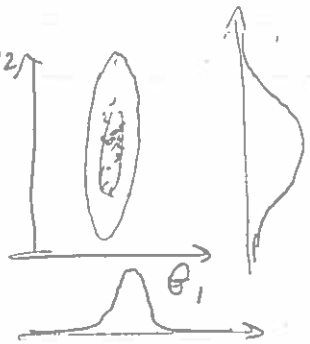
The posterior PDF for one parameter is obtained via marginalization

$$p(\phi | d) = \int \mathcal{L}(\phi, \psi) P(\phi, \psi) \, d\psi$$

Can marginalize over as many dimensions as you want.

If $\mathcal{L}$ can't be computed analytically, then it must be sampled. As long as the samples are drawn with $P \propto \mathcal{L}(\theta)p(\theta)$, the posterior will simply be equal to the histogram of sampled points.

Eg. 2d gaussian




The marginal posterior is found in the same way, by binning in $\theta_1$ and ignoring $\theta_2$

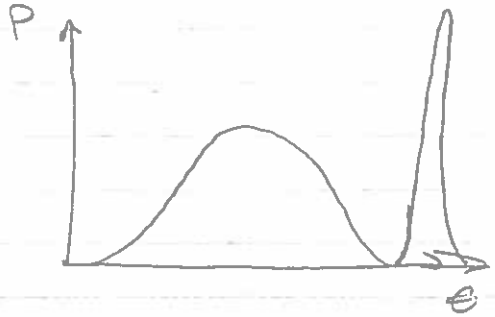
So, even if computing $Z(M) = p(d)$ is not straightforward, the marginal posterior is easy to normalize.

## Reporting

If your posterior distribution is nice and gaussian ← there are weaker conditions.


the point of highest probability (= mode) (posterior max) will converge to the max likelihood (assuming prior not too strong)

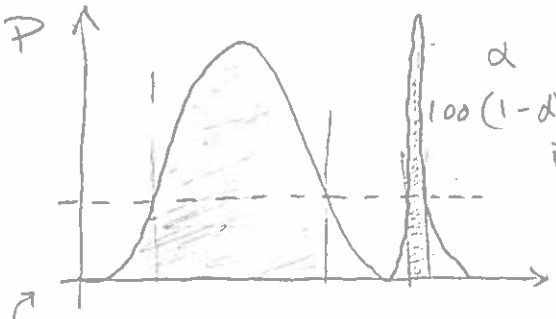
But what if: ("pathological" distr.)




Reporting only $\theta_{max}$ ignores a large amount of posterior probability.

Could report mean, median, but they will also be misleading

We therefore build credibility intervals (or region)

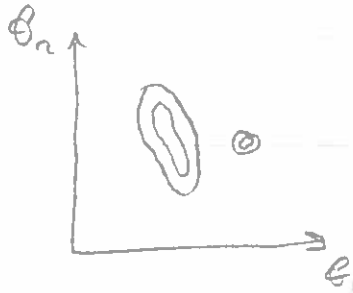



a CI region in which $100(1-\alpha)\%$ of the posterior probabil is contained

Imagine lowering the ---- line until

$$\int_C d\theta\, P(\theta|d) = 1-\alpha$$

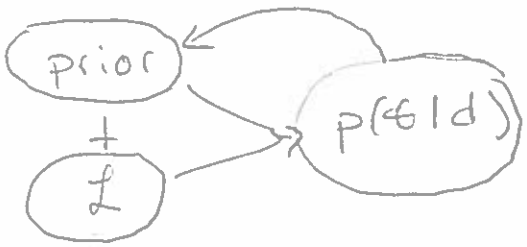
Can do the same in 2d (or 3d if you can)




Note that marginalization takes care of extra volume due to nuisance parameters in a natural way.

If you have a set of samples you can also profile: i.e. find the values of the nuisance parameters that maximize $\mathcal{L}$, and treat the result in a frequentist way.
(Profile Likelihood)

Priors In principle, the posterior from a previous round of experiments become priors for the next




$\Rightarrow$ informative prior

But where do we start if we don't know anything? (this is where people get mad)
We want an uninformative prior.

The most obvious choice is a uniform prior:

$$P(\theta) = \frac{1}{V}$$

This seems great, but there are two potential pitfalls.

1) Concentration of measure: in large D parameter spaces, samples drawn according to flat priors tend to concentrate in a thin shell of constant variance.

2) Reparametrization (non)-invariance. If I want to measure a (unknown) cross section, do I use a uniform prior in $\sigma$? Or $\log\sigma$?

Some jargon:

- Reference priors: prior s.t. contribution of the data to the posterior is maximised.

  A possibility is the "maximum entropy" prior

- Ignorance prior: equal probability to all alternatives. this is not reparam-invariant.

- Conjugate prior: a prior is conjugate if the resulting posterior is same family as as $\mathcal{L}$. Ex: Gaussian is self conjugate. Conj of binomial is $\beta$ fcl.

Reparametrization invariance (RI): The Jeffreys prior is RI : (want $p(\theta)d\theta$ to be invariant)

$$p(\vec{\theta}) \propto \sqrt{\det \mathcal{I}(\vec{\theta})}$$

Where $\mathcal{I}(\vec{\theta})$ is the Fisher information matrix

$$\mathcal{I} = -E\left[\frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(d|\theta)\middle|\theta\right]$$

you can show that $p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right|$.

Maintaing the prior density.

Caveat you need to compute derivatives, which can be expensive if you are sampling.

Note: STAN is a package by Andrew Gelman that apparently computes these.

## Model comparison

Classical hypothesis testing assumes a model is true, and evaluates how unlikely the data are → not $p(H)$, or the relative probability of two models being true.

In Bayesian model selection we use the evidence

$$Z(M) = p(d|M) = \int d\vec{\theta}\, P(d|\vec{\theta})p(\vec{\theta})$$

↓ really $p(\theta|M)$

We can use Bayes' thm again to invert the conditioning:

$$p(\mathcal{M}|d) \propto p(\mathcal{M})\, \underbrace{p(d|\mathcal{M})}_{\mathcal{Z}}$$

$p(\mathcal{M})$ → prior assigned to model itself. (can be $1/N_{\mathcal{M}}$)

("betting odds")

Then the posterior odds are

$$\frac{p(\mathcal{M}_0|d)}{p(\mathcal{M}_1|d)} = B_{01}\, \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}$$

↳ Bayes factor → tells us whether data favour $\mathcal{M}_0$ ($B_{01}>1$) or $\mathcal{M}_1$ ($B_{01}<1$)

These are usually interpreted with the Jeffreys scale:

| $\ln B_{01}$ | odds | prob. | strength |
|---|---|---|---|
| $<1$ | $\lesssim 3:1$ | $<75\%$ | inconclusive |
| 1 | $\sim 3:1$ | 75% | weak evidence |
| 2.5 | 12:1 | 92.3% | moderate " |
| 5 | 150:1 | 99.3% | strong " |

Note that Bayesian model comparison includes an occam's razor effect: extra free params are automatically penalized.

ex: binomial distr (coin toss)

d: 115 heads
85 tails

$$p(d|q) = \binom{200}{85} q^{115} (1-q)^{85}$$

$M_0$: $q = 1/2$

$M_1$: $q$ is a free param $\in [0,1]$

Note that <u>profiling</u> tells us that $M_1$ is a better fit, since the max $\mathcal{L}$ is at $q = \frac{115}{200}$

(usually you would then do a $\chi^2$ per dof...)

In bayesland

$$p(d|M_0) = p(d|q=1/2) = 0.005956$$

$$p(d|M_1) = \int_0^1 dq\, p(d|q) = 0.004975$$

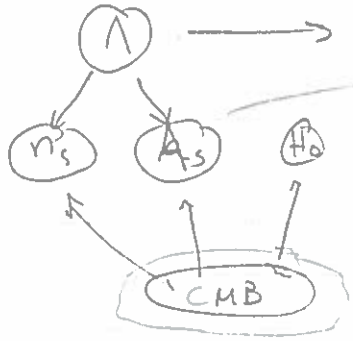
The extra low-probability "volume" of parameter space is weighting down $M_1$

- Note that this says nothing about params that do not affect data

## Hierarchical models

Say your model has a parameter whose prior depends on another → can build a hierarchical model

e.g. I have some parameter $\Lambda$ from an inflation model that affects priors on $A_s$, $n_s$




→ hyperparameter, w/ hyperpriors.

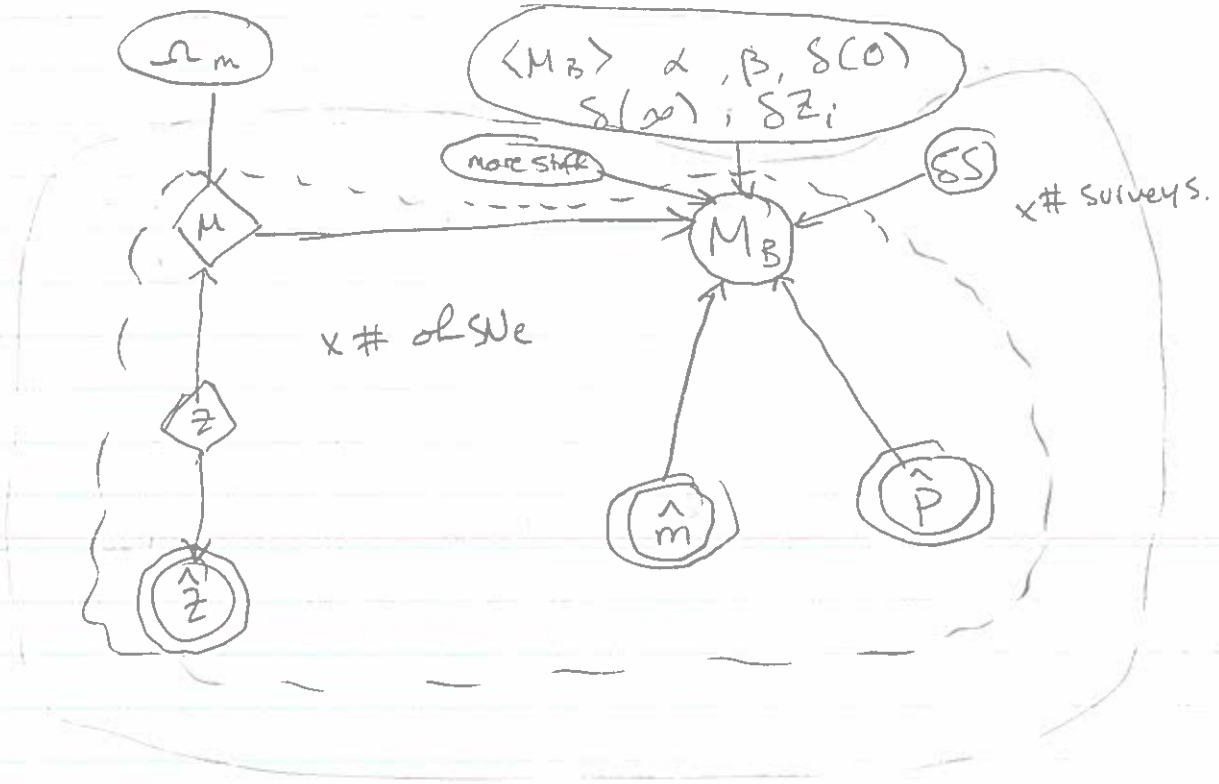
→ prior on $A_s$ could be a gaussian w/ mean $\mu(\Lambda)$ (i.e $p(n_s) = p(n_s|\Lambda)$

Can easily use marginalisation to infer

$$p(A_s) \propto \int d\Lambda\, p(A_s, \Lambda)$$ ← joint posterior

or $$p(\Lambda) \propto \int dA_s\, p(A_s, \Lambda)$$

Very popular in SN:




Steve:

# Sampling

We know that if we have a set of points sampled from $\mathcal{L}(\theta)\,p(\theta)$, their distribution will be proportional to the posterior

Rejection sampling → curse of dimensionality

⇒ MCMC

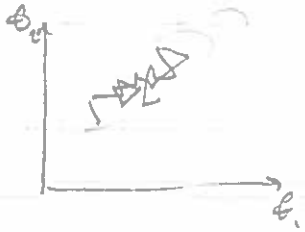
- Monte Carlo = random
- Markov chain → set of successive draws

Can show that an MCMC will converge to a stationary state, with $f(\vec{\theta}) \propto p(\vec{\theta}|d)$ when sampled from the target distribution (same)

Condition to obtain MC: detailed balance

$$p(\theta^{(t)}|d)\,T(\theta^{(t)}, \theta^{(t+1)}) = p(\theta^{(t+1)}|d)\,T(\theta^{(t+1)}, \theta^{(t)})$$

$T$: transition probability




$$p(\theta|d) \propto \text{Hist}(\{\vec{\theta}^{(i)}\})$$

Many ways to ensure detailed balance

Metropolis-Hastings:

i) Start with a random point $\theta^{(0)}$, evaluate its $\mathcal{L}(\theta^{(0)}) = p(d|\theta^{(0)}) \xrightarrow[\text{Bayes}]{} p(\theta^{(0)}|d)$

ii) Draw a candidate next point $\theta^{(c)}$ from a proposal distribution $q(\theta^{(0)}, \theta^{(c)})$, e.g. a gaussian centred at $\theta^{(0)}$

↳ Metropolis algo. $q(\theta^0, \theta^c) = q(\theta^c, \theta^0)$

iii) evaluate $p^c = p(\theta^c|d)$. Accept with prob

$$\alpha = \min\left(\frac{p_c\, q_{c\to 0}}{p_0\, q_{0\to c}}, 1\right)$$

iv) rejection → stay at $\theta^0$ (count that point twice in chain)
otherwise go to $\theta^c$

GOTO ii)

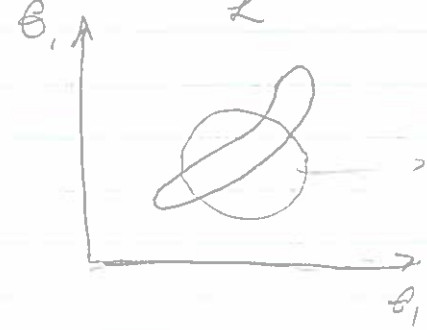
Metropolis: $T_{t\to t+1} = q_{t\to t+1}\,\alpha$ (if $\alpha < 1$)
$T_{t+1\to t} = q_{t\to t+1} \times 1$

⇒ detailed balance is guaranteed.

So if $p_c > p_0$ you always accept the candidate, but sometimes accepting jumps w/ lower $\alpha$ means you don't get stuck.

Optimally, you want an acceptance rate around 25%

Choice of $q$ proposal distr. is very important




a $q$ like this will waste a lot of time sampling in the wrong direction

↳ Soln run a quick MCMC

↓

compute covariance matrix at $p(\theta|y)$

↓

choose $q$ to be Gaussian w/ $\Sigma = Cov(p)$

↓

run Roller MCMC

## Gibbs sampling

Start as before, but for each draw hold every parameter except one fixed (this is "conditional" sampling

This works well in high D space

Block Gibbs sampling is similar

The 1D sampling can be done e.g. by rejection sampling or something more sophisticated.

- Can "block gibbs"
- Can combine w/ MH in fun ways.

## Other useful things

- Hamiltonian sampler
- Importance sampling → can update a chain w/ new information

## Issues w/ MCMC

- Autocorrelated (should thin a chain) but correlated result can often be closer to true target than thinned chain
- Burn-in: initial samples are correlated w/ first guess. → need to discard ~ 10 - 25 % of samples
- Can get "trapped" → "temperature" of proposal is tricky to get right.
- How to assess convergence?

## Nested sampling

It's neat