## Bayesian Methods for Biostatistical Applications

Paul Gustafson
Department of Statistics
University of British Columbia

August 17, 2008

## With whom it all began

Rev. Thomas Bayes, 1702-1761

## The plan for today

- Module #1: Some Bayesian basics (35 minutes?)
- Aside A: Bayesian computation (25 minutes?)
- Module #2: Bayes for imperfect data (40 minutes?)
- Module #3: Bayes for flexibility (30 minutes?)
- Aside B: Bayesian model assessment (20 minutes?)
- Module #4: Some Bayesian subtleties (30 minutes?)

## Module #1

Some Bayesian Basics

## Bayes Theorem

Knowledge about **truth** ($T$) having observed **data** ($D$)?

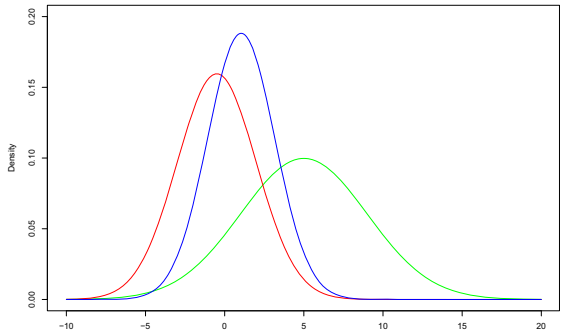
$$Pr(T = t|D = d) = \frac{Pr(D = d|T = t)Pr(T = t)}{\sum_{t^*} Pr(D = d|T = t^*)Pr(T = t^*)}$$

Some explanations are more likely than others!

Or in more parametric ('likelihood $\times$ prior') terms:

$$\pi(\text{parameters}|\text{data}) \propto \pi(\text{data}|\text{parameters})\pi(\text{parameters})$$

## The textbook picture




## Prior distributions: Blessing or curse?

All inferences are reported in terms of state-of-knowledge after the data are observed.

*What do the data add beyond the prior distribution?*

Some controversy still abounds.

**Do the benefits of Bayes lie in:**

- the ability to infuse prior information?
- the principled mechanism to update from prior to posterior?

## Starting from a position of ignorance

'Let the data speak for themselves'

**Formally 'flat' priors**

- $\pi(\mu) \propto 1$
- $\pi(\sigma) \propto \sigma^{-1}$ [or equivalently $\pi(\sigma^2) \propto \sigma^{-2}$]

**Or prior but diffuse priors**

- $\mu \sim N(0, 10^6)$
- $\sigma^2 \sim IG(10^{-4}, 10^{-4})$

Nice frequentist properties sometimes, but need to establish properness and/or reasonableness on a case-by-case basis.

## Actually infuse prior information

'We don't do science in a vacuum.'

### Formal elicitation

- underlying axioms, gambling connections,
- generally regarded as difficult,
- only yields 'my' prior.

### Less formal: the 'as strong as is generally defensible' approach

For instance, say $\beta$ is a (conditional) log odds-ratio between exposure and disease in an epidemiological context.

$\beta \sim N\{0, (0.5 \times \log 8)^2\}$ generally defensible?

## Prior sensitivity

Forwards: report inferences arising from a variety of prior specifications.

Backwards: say the chosen prior distribution yields a 95% interval estimate which exludes zero (i.e., a 'significant' finding).

- Determine how much this prior would have to be perturbed in order to lose this significance.
- Comment on the plausibility of this perturbed prior.

## Bayesian computation (come back to this later)

### Ideal

(model specification, prior specification, data)

$\downarrow$

'black box'

$\downarrow$

(exact representation of) posterior distribution

## Computation, continued

### 2008 Reality

- Markov chain Monte Carlo (MCMC) methods: represent the posterior distribution approximately via an (arbitrarily large) sample simulated (with caveats) from it.
- Colour of box? Anywhere from white (C/Fortran) to dark-grey (WinBUGS).

Will return to this.

## Inferential summaries (and how obtained via MCMC)

$\pi(\theta|\text{data})$: joint posterior distribution over all parameters.

$\pi(\beta|\text{data})$: **marginal posterior distribution** of scalar $\beta = g(\theta)$, describing post-study knowledge of $\beta$.

(Histogram or kernel density estimate of MCMC-sampled $\beta$ values.)

**Point estimate** of $\beta$ - mean/median/mode of $\pi(\beta|\text{data})$.

(mean/median/mode of MCMC-sampled $\beta$ values)

Decision-theory connection:
which loss function implies which point estimator.

## Inferential summaries, continued

**95% interval estimate:**

**Equal-tailed:** (2.5, 97.5)-th percentiles of $\pi(\beta|\text{data})$.

(percentiles of MCMC-sampled $\beta$ values)

**Highest-posterior-density (HPD):** - shortest interval containing 95% probability under $\pi(\beta|\text{data})$.

(search over $(a, 95 + a)$-th percentiles of MCMC-sampled $\beta$ values)

**Predictive distribution:**

$$\pi(\text{datum}_{n+1}|data_1^n) = \int \pi(\text{datum}_{n+1}|\theta)\pi(\theta|\text{data}_1^n)d\theta$$

(augment MCMC-sample with draws from $[\text{datum}_{n+1}|\theta]$)

## Illustration: Misclassified exposure

### Variables

- $Y$ binary outcome
- $X$ actual binary exposure
- $X^*$ possibly misclassified binary exposure

Interested in $(Y, X)$ association, but have $(Y, X^*)$ data.

### Misclassification parameters

Sensitivity: $SN = Pr(X^* = 1|X = 1)$
Specificity: $SP = Pr(X^* = 0|X = 0)$.

## Type of misclassification

Also important: whether the misclassification is nondifferential (blind to outcome) or differential.

Do $Pr(X^* = 1|X = 1, Y = y)$ and $Pr(X^* = 0|X = 0, Y = y)$ vary with $y$?

Often an issue in case-control designs, for instance, especially with self-report of exposure status.

For the sake of a simple illustration right now, we assume nondifferential misclassification.

## Probability model (likelihood)

Consider case-control design.

**Actual exposure prevalences:** $r_i = Pr(X = 1|Y = i)$, for $i = 0$ (controls) and $i = 1$ (cases).

Summary statistics: $Z_0$ out of $n_0$ controls and $Z_1$ out of $n_1$ cases are **apparently exposed** ($X^* = 1$).

$$Z_i \sim Bin\{n_i, r_i SN + (1 - r_i)(1 - SP)\},$$

for $i = 0, 1$.

Inferential target is **log-odds-ratio**: logit $r_1$ − logit $r_0$.

## Some prior specifications

Aim for 'epidemiologically defensible' prior distribution for $(r_0, r_1)$:

$$\begin{pmatrix} \text{logit}r_0 \\ \text{logit}r_1 \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu \\ \mu \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}.$$

For instance, choose $(\mu, \tau, \rho)$ such that 95% prior probability of:

- $p_{lo} < r_i < p_{hi}$,
- $\log(1/k) < \text{logit}r_1 - \text{logit}r_0 < \log(k)$.

E.g., investigator selects $p_{lo} = 0.02$, $p_{hi} = 0.50$, $k = 8$.

## Some prior specifications, continued

Contrast three prior specifications for $(SN, SP)$.

1. $SN \equiv 1$, $SP \equiv 1$:
   pretend there is no misclassification.
2. $SN \equiv$???, $SP \equiv$???:
   admit there is misclassification, but assume its magnitude is known exactly.
3. $SN \sim Beta(???, ???)$, $SP \sim Beta(???, ???)$
   admit there is misclassification, with magnitude known only approximately.

Helpful to think of all three as 'priors,' even though the first two are not stochastic. In particular, 1 and 2 are not more objective than 3.

## A glimpse at the innards: WinBUGS model specification

```
model {
for (i in 1:2) {
  z[i] ~ dbin(p[i], n[i])
  p[i] <- sens*r[i] + (1-spec)*(1-r[i])
  logit(r[i]) <- rr[i]
  rr[i] ~ dnorm(rstr, r1prec)
}

sens ~ dbeta(a.sn, b.sn); spec ~ dbeta(a.sp,b.sp)

rstr ~ dnorm(mu, r2prec)
r1prec <- 1/((1-rho)*tau2); r2prec <- 1/(rho*tau2)

rho <- 1 - pow(log(k),2)/(8*tau2)
tau2 <- pow( (logit(p.hi)-logit(p.lo))/4, 2)
mu <- (logit(p.hi)+logit(p.lo))/2
}
```

## Example

Data on maternal use of antibiotics during pregnancy and sudden infant death syndrome (SIDS).

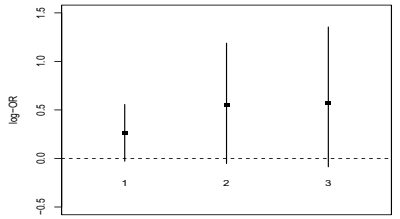
**Apparent exposure** is self-report of antibiotic use on questionnaire.

101 of 580 controls and 122 of 564 cases apparently exposed.

Validation study suggests $SN \approx 0.6$, $SP \approx 0.9$.

In fact,
$SN \sim 0.60 \pm 0.10 \rightarrow$ Beta(57, 38) prior
$SP \sim 0.90 \pm 0.03 \rightarrow$ Beta(359, 40) prior

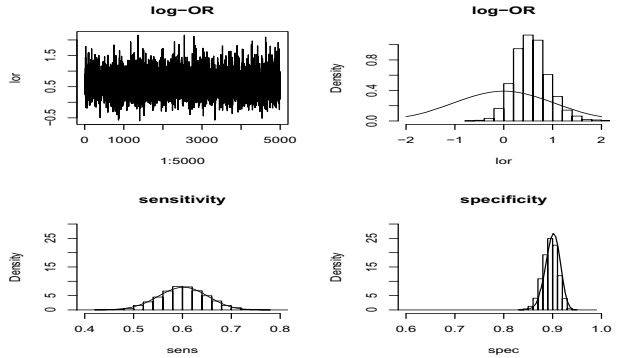
## Estimated log-OR under the three specifications




Admitting to misclassification pushes point estimate (but not lower limit of 95% interval estimate) away from the null.

Admitting to uncertainty about the magnitude of misclassification further increases uncertainty, albeit modestly.

## Some details from third specification




## Some references

### General Bayesian books

- Gelman *et. al.* (2nd ed., 2004)
- Carlin and Louis *et. al.* (3rd ed., 2008)
- Berger (2nd. ed. 1985)
- And many others: Gill, Congdon, Robert, Moyé,...

Misclassification ex.: Gustafson, Le, and Saskin (2001, Biometrics)

## ASIDE A

Bayesian Computation

## MCMC

Regardless of platform (C/R/WinBUGS), the cornerstone idea of Markov Chain Monte Carlo (MCMC) is as follows.

Would like to 'know' the distribution having density proportional to $f(z)$, i.e., graphically represent its marginal distributions, numerically represent its moments, quantiles, etc.

Obvious application: $f()$ is unnormalized posterior density, i.e., likelihood times prior.

A large computer-simulated sample of realizations from $f()$ suffices to 'know' $f()$ to good approximation.

Note the nested use of statistical thinking!

## MCMC, continued

**Given the (unnormalized) target $f(z)$, construct a Markov Chain $Z^{(0)}, Z^{(1)}, \ldots$ having $f(z)$ as its stationary distribution.**

- Metropolis-Hastings algorithm is a very general method of constructing such a chain.
- Theory indicates that for any choice of dist $\{Z^{(0)}\}$, the chain will converge to its stationary distribution.
- Hope that the **pleasantness of the target** plus our **clever choice of particular chain construction** is such that:
  - the convergence is quick, i.e., dist$(Z^{(b)})$ is very close to $f()$ for relatively small $b$.
  - the chain mixes well, i.e., dist$(Z^{(i)}, Z^{(i+1)})$ not too small, on average.

## Metropolis-Hastings algorithm

Have $Z^{(i)}$

Generate $Z^*$ according to $t(Z^{(i)} \to Z^*)$

Compute

$$p = \min\left\{\frac{f(Z^*)t(Z^* \to Z^{(i)})}{f(Z^{(i)})t(Z^{(i)} \to Z^*)}, 1\right\}.$$

Set

$$Z^{(i+1)} \leftarrow \begin{cases} Z^* & \text{with probability } p, \\ Z^{(i)} & \text{with probability } 1-p. \end{cases}$$

## MCMC, continued

We can treat the 'post-burn-in' output $Z^{(b+1)}, \ldots, Z^{(b+m)}$ as a *did* (dependent and identically distributed) sample from the target distribution.

i.e., If we thinned this sample more and more aggressively, it would look more and more like an *iid* sample from the target.

i.e., The *did* sample of size $m$ is as good at summarizing the target distribution as an *iid* sample of size $m^*$, where $m^* << m$.

(But this is a conceptual argument – shouldn't actually thin unless storage issues demand it!)

## MCMC: chain construction

Often the best strategy is to construct the chain from univariate Metropolis-Hastings transitions:

| update | target density |
|---|---|
| $\left\{Z_1^{(i+1)} \mid Z_2^{(i)}, Z_3^{(i)}\right\}$ | $f(z_1 \mid z_2, z_3)$ |
| $\left\{Z_2^{(i+1)} \mid Z_1^{(i+1)}, Z_3^{(i)}\right\}$ | $f(z_2 \mid z_1, z_3)$ |
| $\left\{Z_3^{(i+1)} \mid Z_1^{(i+1)}, Z_2^{(i+1)}\right\}$ | $f(z_3 \mid z_1, z_2)$ |

i.e., the $j$-th '**full-conditional distribution**' under $f()$ is the target density when updating $Z_j$.

Best case: each full conditional is a standard distribution.
Simply sample each full conditional in turn — the Gibbs sampler.

Typical case: at least some full conditionals are 'messy.'
Turn to 'random-walk' Metropolis-Hastings and other algorithms.

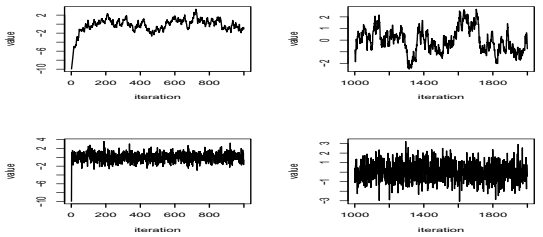
## Chain construction, continued

Generally, spend one's time:

- debugging,
- waiting for the computer to finish,
- assessing burn-in and convergence.

## MCMC: traceplots

More formal assessment possible and desirable, but as a first step look at traceplots.




Dislike the top chain, like the bottom one - confident it is a numerically accurate representation of the target distribution.

## MCMC: traceplots, continued

Actually, a minor confession:

The second chain is a thinned version (every 100-th iteration) of the first.

In the face of plentiful computing resources and some patience, it may not be worth trying to devise a great sampler. A passable sampler with long runs will look/be fine.

## A few comments on WinBUGS ... from a recent convert ...

Learning curve: less than one day.

Unless the problem you face is hugely sophisticated, worth trying WinBUGS first, because feasibilty or lack thereof will quickly become apparent.

Can always later turn to:
R: to control which algorithm updates which parameter,
C: for fast/scalable code.

If (like me) you are less enamoured with GUIs and menus, consider calling WinBUGS from within R using the R2WinBUGS package: see Andrew Gelman's homepage.

## Some references

The general Bayes books cited earlier.

Many WinBUGS examples on web.

### Specific MCMC books

- Robert and Casella (2004, 2nd. ed.)
- Liu (2001)
- Chen, Shao and Ibrahim (2001)
- Gilks et. al. (Eds.) (1996)

## MODULE #2

Bayes for Imperfect Data Problems

## Collapsed and Augmented Views

In the misclassification example, we analytically determined $(X^*|Y)$, proceeded with:

$$\pi(r_0, r_1|x^*_{1:n}, y_{1:n}) \propto \prod_{i=1}^{n} \pi(x^*_i|y_i, r_0, r_1, sn, sp) \times \pi(r_0, r_1, sn, sp),$$

the collapsed view.

But, could also have thought in terms of:

$$\pi(r_0, r_1, sn, sp, x_{1:n}|x^*_{1:n}, y_{1:n}) \propto \prod_{i=1}^{n} \pi(x^*_i|x_i, sn, sp)\pi(x_i|y_i, r_0, r_1) \times \pi(r_0, r_1, sn, sp),$$

the augmented view.

## Augmented view more generally

Ideal but unobservable data $D^{(I)}$.

Imperfect but observable data $D^{(O)}$.

$$\pi(\theta, D^{(I)}|D^{(O)}) \propto \pi(D^{(O)}|D^{(I)}, \theta)\pi(D^{(I)}|\theta)\pi(\theta).$$

MCMC algorithms inherit the flavour of repeatedly:

- imputing $D^{(I)}$ given $D^{(O)}$ and $\theta$,
- imputing $\theta$ given $D^{(I)}$ (and possibly $D^{(O)}$).

## Why take augmented view

- **PRO:** applicable in many problems where $D^{(O)}|\theta$ has no analytical form with which to tackle $\theta|D^{(O)}$ directly.
- **PRO:** Very typically MCMC updates to $\theta$ are nicer when conditioning on $D^{(I)}$ than when conditioning on $D^{(O)}$.
  - for instance, Gibbs-sampling updates (no tuning) versus RWMH updates (require tuning).
- **PRO:** allows individual level inference on $D^{(I)}$, if desired.
- **CON:** turns a $p$-dimensional integration problem into a $(p + n)$-dimensional integration problem.

## Why take augmented view, continued

The augmented view tends to prevail, though the choice is not so clear cut, in my view.

In at least one class of problems I've encountered:

- $D^{(O)}|\theta$ has no analytical form,
- $\theta|D^{(I)}$ gives nice MCMC updates.

Yet:

- the augmented approach doesn't work (poor MCMC convergence/mixing)
- using numerical quadrature to evaluate the $\pi(D^{(O)}|\theta)$ inside an MCMC algorithm applied in the collapsed view works OK.

More thought needed?

## Illustration: Framingham heart study

**Cohort 55 or younger at initial exam, n=4526**

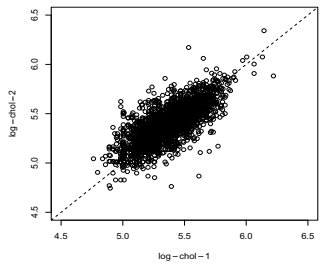
| | |
|---|---|
| $Y$ | **twenty-year** mortality, all cause; |
| $X$ | (true) log serum-cholesterol (log mg/100ml); |
| $X_1^*$ | $X$ as measured at initial exam; |
| $X_2^*$ | $X$ as measured at first follow-up exam; |
| $Z$ | AGE and AGE$^2$, GENDER, SMOKING, REL-WHT, DBP, SBP. |

Want to infer $(Y|X,Z)$ relationship from $(Y, X_1^*, X_2^*, Z)$ data.

## Model Specification

$$\underbrace{\pi(X_1^*, X_2^*|Y, X, Z)}_{\text{measurement model}} \times \underbrace{\pi(Y|X,Z)}_{\text{outcome model}} \times \underbrace{\pi(X|Z)}_{\text{exposure model}}$$

## Measurement model




Try **nondifferential** measurement model for $(X_1^*, X_2^*|X)$ with:
- $X_1^*, X_2^*$ conditionally independent given $X$,
- $(X_i^*|X) \sim N(X, \sigma_e^2)$, $i = 1, 2$.

Corresponds to a multiplicative measurement error for CHOL.

## Outcome and exposure models

- Logistic regression outcome model for $(Y|X,Z)$.
- Normal/linear regression exposure model for $(X|Z)$.
  - Evidence for this?
  - Damage done if wrong?
    (parameters in this model are not of direct interest)

Bear in mind that all three models involve the unobservable $X$, hence residual plots etc., are not (easily) forthcoming.

**Prior distributions:**
- N(0, 'big') for regression coefficients,
- 'Unit-information' priors with conservative guesses for variances.

(So why bother with Bayes???)

## MCMC approach

Cycle through updates to

- $(X_1, \ldots, X_n)$,
- measurement model parameters
- outcome model parameters
- exposure model parameters

Some Gibbs sampling updates, some random-walk MH (RWMH) updates.

Outcome model coefficients:
- mix of fast/lousy updates (RWMH)
- slow/good updates (based on logistic regression ML fit)

## Bayes-MCMC results

'Examination of a posterior sample (of all params. in the three models plus $X$ values for each subject) of size 5000 reveals fast convergence and good mixing'

Inferences for $(Y|X, Z)$ coefficients: posterior mean and SD (as point estimate and 'standard error'):

| | $\hat{\beta}$ | (PSD) | $\exp\left(\hat{\beta}\right)$ |
|---|---|---|---|
| LOG-CHOL | 0.131 | (0.120) | 1.14 |
| $AGE_1$ | 0.083 | (0.008) | 1.09 |
| $AGE_2$ | -0.001 | (0.001) | 1.00 |
| GENDER | -0.475 | (0.091) | 0.62 |
| SMOKING | 0.756 | (0.095) | 2.13 |
| ... | ... | ... | ... |

Coded wrt a change of $\log 1.5 \approx 0.4$ in LOG-CHOL, i.e., a 50% increase in CHOL.

## Lessons learned

| | $\hat{\beta}$ | PSD |
|---|---|---|
| answer adjusting for measurement error: | 0.131 | (0.120) |
| ignoring (taking $X = X_1^*$): | 0.120 | (0.093) |

Yawn. We have 'un-attenuated' to a rather modest extent.

Okay. But hard to determine this without doing the analysis, particularly given the measurement error magnitude.

Also interesting: **evidence for a positive association** between cholesterol and mortality.

| | |
|---|---|
| Ignoring measurement error: | $Pr(\beta > 0 \mid Data) = 0.89$ |
| Adjusting | $Pr(\beta > 0 \mid Data) = 0.86$ |

Ignoring measurement error NOT conservative in this sense!

## Illustration: Unmeasured confounding

**Variables**

- $Y$ disease outcome,
- $X$ exposure,
- $C$ potential confounders that are identified/available.

But concern about $U$, one (or more) confounders that we can't identify or isn't available.

Regular analysis: Regression of $Y$ on $(X, C)$.

Ideal but impossible analysis: Regression of $Y$ on $(X, C, U)$.

Doing the regular analysis and interpreting the $X$ coefficient as 'the exposure effect' effectively corresponds to using a very special prior distribution: with 100% prior certainty, $U$ is not a confounder.

## Unmeasured confounder, continued

Recall $U$ not a confounder if it doesn't drive the outcome or isn't associated with exposure.

**More precisely:**

- $Y$ and $U$ conditionally independent given $(X, C)$,

and/or

- $U$ and $X$ conditionally independent given $C$.

Is this a realistic prior distribution in most applications?

**More realistically:**

Any dependence between $Y$ and $U$ given $(X, C)$ is likely to be limited, as is dependence between $U$ and $X$ given $C$.

Can build a prior of this form.

## Beta-blockers and mortality after heart failure

**n=6969 BC residents discharged alive from hospital with primary diagnosis of heart failure**

- $X$: dispensed beta-blocker within 30 days
- $Y$: one-year all-cause mortality
- $C$: 21 potential confounders (demographics, comorbidities, hospitalization characteristics, HF meds)

Admin. data - concern disease severity ($U$?) not well captured

## BB-MAHF, continued

**Model Structure**

$$\begin{aligned} \text{logit}(Y|X,U,C) &= \alpha + \theta X + \lambda_0 U + \lambda_1 C_1 + \ldots + \lambda_p C_p \\ \text{logit}(U|X,C) &= \omega + \gamma_0 X \\ \text{logit}(X|C) &= \xi + \gamma_1 C_1 + \ldots + \gamma_p C_p \end{aligned}$$

describes ideal (i.e., including $U$) data in terms of parameters.

Specification completed with *prior distributions* for all parameters.

Plausible ranges for strengths of $(U, X)$ and $(U, Y)$ relationships required.

## Prior structure

**Parameters:**

- $\gamma$'s describe association of exposure $X$ with confounders $(U, C_1, \ldots, C_p)$.
- $\lambda$'s describe association of $Y$ with $(U, C_1, \ldots, C_p)$, (given $X$).
- These parameters all conditional log-odds ratios.
- For example, assign $0 \pm log6$ prior in each case.

Different styles of prior: independent versus exchangeable.

## Different styles of prior

**Prior on $\gamma_0, \gamma_1, \ldots, \gamma_p$ (similarly on $\lambda_0, \lambda_1, \ldots, \lambda_p$)**

- Independent and identically distributed.
- Conditionally independent with mean zero and unknown variance (which itself is assigned a prior distribution), hence exchangeable.

Either way, can assign each $\gamma_j$ the same *marginal* distribution (e.g., Student's t, with $df = 10$, $\mu = 0$, $\sigma = 0.5 \times log6$).

Differing implications:
knowledge of $\gamma_j$ implies something about $\gamma_k$?

## Exchangeable prior in action

- Data directly inform values of $\gamma_1, \ldots, \gamma_p$ - how associated is $X$ with $C_1, \ldots, C_p$.
- Thus some indirect information flow to $\gamma_0$, describing the $(X, U)$ relationship.
- Reflects notion that $|\gamma_0|$ is more likely to be small if $|\gamma_1|, \ldots, |\gamma_p|$ tend to be small.
- Similarly for $(\lambda_0, \lambda_1, \ldots, \lambda_p)$ describing associations with $Y$.
- Matches epidemiological folklore???
  If adjusting for a series of potential confounders has little impact, then adjusting for one more is less likely to have a large impact?

## BB-MAHF results

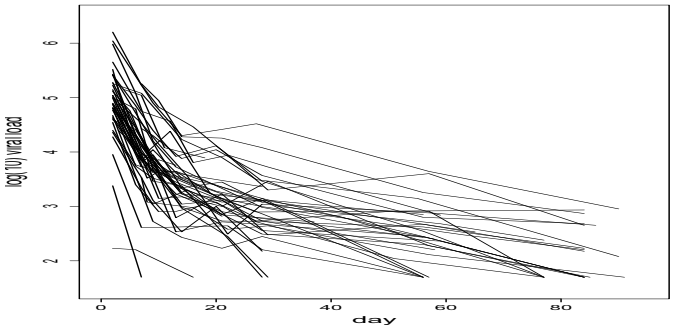
| | | $\hat{OR}$ | |
|---|---|---|---|
| Crude | $(Y\|X)$ | 0.63 | (0.55, 0.74) |
| Naive | $(Y\|X, C)$ | 0.72 | (0.62, 0.86) |
| Ind. Prior | $(Y\|X, U, C)$ | 0.72 | (0.45, 1.15) |
| Exch. Prior | $(Y\|X, U, C)$ | 0.72 | (0.56, 0.93) |

**Simultaneously acknowledge:**

- sampling variability ($n = 6969$ not $n = \infty$),
- unknown relationship between $U$ and $(X, Y, C)$.

## Illustration: Viral load modelling

Viral load (HIV) over time, $n = 44$ treated patients.
Wu & Ding (1999), Ko & Davidian (2000), Wu (2002).



## Outcome model

$Y_{it}$ $\log_{10}$ viral load of patient $i$ at time $t$;
$X_i$ baseline (CD4 cell count) - poorly measured;
$Z_i$ baseline (log) viral load.

Complex nonlinear model to explain viral dynamics (initial/decay phases, random effects):

$$Y_{it} \sim N\left(\theta_{1i}e^{-\lambda_{1i}t} + \theta_{2i}e^{-\lambda_{2i}t}, \sigma^2\right) \quad (\lambda_{1i} > \lambda_{2i})$$

$$\begin{aligned} \log\theta_{1i} &= \beta_1 + \beta_2 Z_i + b_{1i} \\ \log\theta_{2i} &= \beta_3 + \beta_4 Z_i + b_{2i} \\ \log\lambda_{1i} &= \beta_5 + \beta_6 X_i + b_{3i} \\ \log\lambda_{2i} &= \beta_7 + b_{4i} \end{aligned}$$

Modelling impact of baseline CD4 count on initial decline.

## Exposure and measurement models

Normal, linear exposure model for $(X|Z)$.

Simplest possible (nondifferential) measurement model:

$$X^*|X, Z, Y \sim N(X, \sigma_e^2).$$

Knowledge of $\sigma_e$? No validation sub-sample, no replicates.

Ko & Davidian (2000), guided by subject-area knowledge, sensitivity analysis: try fixed values $\sigma_e^2 = 0, 0.12, 0.24, 0.36$.

Or maybe one should use this guidance to form an 'informative' prior distribution for $\sigma_e^2$ ... one day ....

## Priors and MCMC

Combination of 'flat' arguments, plus 'best guess with lots of uncertainty' arguments to assign prior distributions.

MCMC in augmented view: unobserved $x_{1:n}$, random effects $b_{j,1:n}$, $j = 1, \ldots, 4$, corresponding variance components, coefficients $\beta_{1:7}$.

So cycle through $5n + 11 = 231$ univariate updates to do one overall MCMC update.

Slow-mixing, fiddly, hard-to-tune, but tolerable.

## Results

Recall: $\beta_6$ governs association between baseline CD4 count and initial decline in viral load.

Recall: $\sigma_e^2$ describes measurement error in baseline CD4 count.

| | $\sigma_e^2$ | 0 | 0.12 | 0.24 | 0.36 |
|---|---|---|---|---|---|
| (posterior mean) | $\hat{\beta}_6$ | 0.122 | 0.191 | 0.448 | 0.642 |
| (posterior SD) | "SE" | 0.089 | 0.135 | 0.236 | 0.287 |

Typical: correcting for bias toward null, loss of info.

**Why am I showing you this example?**

Proof of concept: we can do hard problems, with some effort.

## Some references

The three examples

- Gustafson (2003, Ch. 4.3 and Ch. 4.5)
- McCandless, Gustafson and Levy (*Stat. Med.* 2007, *J. Clin. Epi.* 2008)

Recent book on Bayes-MCMC approaches to missing data: Daniels and Hogan (2008)

## Module #3

BAYES FOR FLEXIBILITY

## Two (at least) kinds of flexibility

**Flexible modelling in terms of:**

1. 'functional forms' used,
2. 'structural assumptions' used.

## 'Functional form' flexibility

Avoiding restrictive distributional assumptions.

- Bayesian nonparametrics
  - Dirichlet processes, Polya trees, etc.
  - burdgeoning literature
  - amazing technical prowess
  - not immodest learning curve
- Bayesian 'mid-to-many' parameters
  - parsimony-encouraging priors
  - hierarchical, extent of encouragement not fixed

## Functional form flexibility, textbook example

Replace

$$E(Y|X) = \beta_0 + \beta_1 X$$

with

$$E(Y|X) = \sum_{i=1}^{p} \alpha_i B_i(X)$$

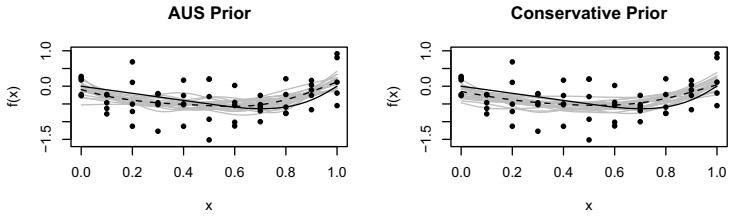
E.g., $\{B_1(x), \ldots, B_p(x)\} = \{1, x, x^2, x^3, (x - c_1)_+^3, \ldots, (x - c_p)_+^3\}$.
Hierarchical prior on $\alpha$ such that linearity encouraged,
to an unknown extent, e.g.,

$$\begin{aligned} \alpha|\tau^2 &\sim N_k(\mu, \tau^2 M) \\ \tau^2 &\sim ??? \end{aligned}$$

where $\alpha^T M^{-1} \alpha$ measures curvature of $E(Y|X)$.

## (Partial) Illustration

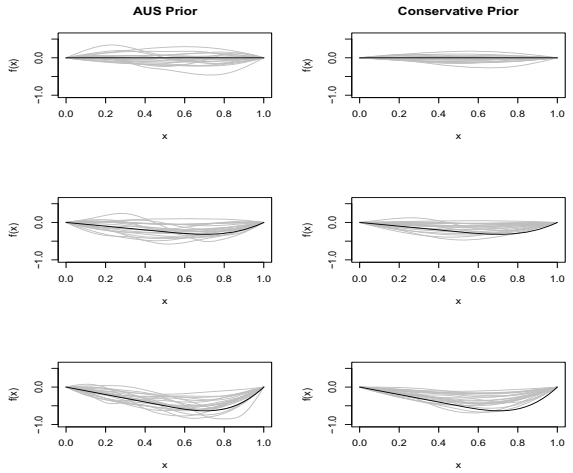
'Default' and conservative prior distributions for $\tau^2$




True curve, posterior mean, posterior draws

## (Partial) Illustration, continued

True curves and posterior means under repeated data-generation.




## Philosophical note

To me, this example epitomizes some advantages of the Bayesian approach.

Take advantage of prior knowledge at a meta-level,
i.e., a smoother curve is a more plausible than a rougher one for most phenomena,
rather than at a detail-level,
i.e., the smoothness of the phenomenon in question is likely between 5.6 and 17.2.

And the answer (posterior distribution of the regression function) automatically incorporate uncertainty about the smoothing parameter ($\tau^2$) - compare with 'cross-validate and plug-in' approaches which are common in smoothing problems.

## A word more on the "conservative" prior

General set-up of linear mixed model
(curve-fitting, spatial-data, multicentre-study data, etc.).

Set-up to guard against **substantial overestimation of random-effect variances**.

Functional form of $\pi_{CONS}(\tau^2)$ somewhat involved in general, depends on design matrix and covariance structure for the random effects.

In simple problems, looks like

$$\pi(\tau^2|\sigma^2) \propto \frac{1}{\sigma^2(1+\tau^2/\sigma^2)^{a+1}}.$$

## Another example: skew extensions of normal family

| | density |
|---|---|
| normal | $\phi(z)$ |
| skew-normal | $\phi(z)2\Phi(\omega_1 z)$ |
| generalized-skew-normal | $\phi(z)2\Phi(\omega_1 z+\omega_3 z^3)$ |
| generalized-skew-t | scale mixture thereof |

Also note that $X \sim GST(\mu,\sigma^2,\omega_1,\omega_2,\mathrm{df})$ can be expressed as:

$$\begin{aligned} X|S &\sim GSN(\mu,\sigma^2S^2,\omega_1,\omega_2) \\ S^2 &\sim Gamma(\mathrm{df}/2,\mathrm{df}/2) \end{aligned}$$

Suggests 'expanded-view' MCMC strategy, applied to $(S_1,\ldots,S_n,\mu,\sigma,\omega_1,\omega_3)$.

## Application to measurement error models

| | |
|---|---|
| measurement model | $X^*\|Y,X,Z$ |
| outcome model | $Y\|X,Z$ |
| exposure model | $X\|Z$ |

Use of $N(\alpha+\beta Z,\sigma^2)$ exposure model often cited as a criticism.

Move to $GST(\alpha+\beta Z,\sigma^2,\omega_1,\omega_3,\mathrm{df})$ exposure model.

Center prior at $\omega_1=0$, $\omega_3=0$, $\mathrm{df}=\infty$.

Hossain (2007 Ph.D. thesis): for all true exposure distributions tested (including some non-GST), removes virtually all bias in estimating outcome model regression coefficients.

Also performs well in comparison to so-called *functional methods* which attempt to avoid any specification of an exposure model.

## Flexibility about structural assumptions

Analysis of imperfect observational data can involve tension surrounding conditional independence assumptions which are:

- needed to ensure the data are informative about the quantity of interest,
- dubious, and empirically uncheckable due to data imperfections.

Bayesian analysis allows the use of a compromise:
Specify a prior such that large departures from conditional independence are unlikely.

Non-Bayesian alternatives to this?

## Illustration

SIDS case-control study again:

unobserved $X$: maternal anemia during pregnancy
observed $X_1^*$: $X$ measurement via questionnaire
observed $X_2^*$: $X$ measurement via chart review

| | controls | | cases | |
|---|---|---|---|---|
| | $X_2^*=0$ | $X_2^*=1$ | $X_2^*=0$ | $X_2^*=1$ |
| $X_1^*=0$ | 147 | 15 | 125 | 15 |
| $X_1^*=1$ | 34 | 20 | 49 | 24 |

Note: a saturated model would have six parameters.

## Illustration, continued

Model for $X|Y$:

- two independent binomial counts,
- log-OR is target parameter.

Possible models for $X_1^*, X_2^*|X, Y$:

1. discard discordant pairs, take $X$ to be common value,
2. conditional independence, $(SN_i, SP_i)$, $i = 1, 2$,
3. prior which relaxes conditional independence assumption

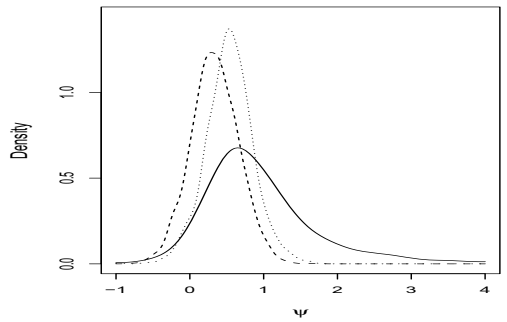
Note on Model 2:

- saturated.

Notes on Model 3:

- truncated exponential priors on $\mathrm{Cov}(X_1^*, X_2^*|X)$,
- downweight corr=0.25 by a factor of four relative to corr=0.

## Illustration, continued: Posterior distribution of log-OR




1-dashed (concordant)
2-dotted (conditional independence)
3-solid (conditional independence relaxed)

## Some references

Bayesian curve-fitting (Dennison *et. al.* 2002).

(Choice of prior: Daniels 1999, Gustafson, Hossain and MacNab 2006, *CJS*).

Bayesian nonparametrics: many people, look at David Dunson's work in particular for leading-edge methods applied to biostatistical problems.

FGST in measurement error models:
Hossain, 2007 Ph.D. thesis, UBC.

Relaxing conditional independence assumptions:
Gustafson (2003, Ch. 5.3, 2005 *Stat. Sci.*, *Stat. Med.*, 2007 *JRSS-B*).

## Aside B

**MODEL SELECTION / COMPARISON / AVERAGING**

Some passing thoughts from a non-expert.

## Principled Bayesian model assessment choice or averaging!)

A single model-prior specification yields a joint distribution of data and parameters.

Competing model-prior pairs: $\pi_a(\text{data}, \theta_a)$ and $\pi_b(\text{data}, \theta_b)$.

Indicator $M$ of true model - just another parameter under the Bayesian lens - requires a prior.

$$\begin{aligned} Pr(M = a|\text{data}) &= \frac{Pr(\text{data}|M = a)Pr(M = a)}{\sum_{m\in\{a,b\}} Pr(\text{data}|M = m)Pr(M = m)} \\ &= \frac{\pi_a(\text{data})Pr(M = a)}{\sum_{m\in\{a,b\}} \pi_m(\text{data})Pr(M = m)} \end{aligned}$$

where 'the marginal' for a given model is:

$$\pi_m(\text{data}) = \int \pi_m(\text{data}, \theta_m)d\theta_m = \int \pi_m(\text{data}|\theta_m)\pi_m(\theta_m)d\theta_m.$$

## Purported 'three strikes' against principled Bayesian model assessment

1. Hard to compute - doesn't 'fall out' of MCMC output for individual models, while MCMC for the 'uber-model' can be challenging.
2. Answers can be sensitive to choices of within-model priors $\pi_a(\theta_a)$, $\pi_b(\theta_b)$.
3. Seemingly amorphous sense of 'best model' compared to very empirical criteria based on cross-validation, penalized likelihood, etc.

## Strike #1: Computation

Indeed a big issue. Many years of research on:

- Add-ons to compute $\pi_m(\text{data})$ given MCMC output on $\pi_m(\theta_m|\text{data})$,
  - some promise (see, for instance, work of Chib), but not yet black-box-ized, i.e., not in WinBUGS.
- Schemes to do MCMC directly on $(M, \theta_a, \theta_b|\text{data})$, e.g., reversible-jump MCMC (Green, 1995).
  - some successes,
  - some tuning angst,
  - also not yet black-box-ized.

## Strike #2: Sensitivity to within-model priors

In many situations, the posterior over the model space, $\pi(M|\text{data})$ is indeed quite sensitive to the choice of within-model priors, $\pi_a(\theta_a)$ and $\pi_b(\theta_b)$, even for fixed across-model prior $\pi(M)$.

Regard as weakness of formal Bayesian model assessment?

I regard it as a statement of 'fundamental sensitivity' whenever model assessment/selection is found.

- flatish objective functions when choosing smoothing parameters via cross-validation,
- big variation when most selection techniques (e.g. stepwise regression) are applied to repeated bootstrap samples,
- arguments, and different answers, concerning ?IC (AIC, BIC, FIC, GAIC, etc.).

## Strike #3: Preference for something more empirical

Some are happy with very empirically motivated schemes to select models, i.e., explicitly based on how well the model does when parts of the data are used to predict other parts of the data.

For instance, pick the model maximizing $\sum_i \log \pi(x_i|x_{-i})$.

Or, instead of emphasizing $(n-1) \to 1$ predictions, aggregate predictions of all sizes. For instance, pick the model maximizing $\sum_i \log f(x_i|x_{1:(i-1)})$.

Wait a minute - this is formal Bayesian model assessment — choosing the model-prior pair yielding the largest marginal density evaluated at the observed data. Pretty empirical after all!

## Deviance Information Criterion (DIC)

Becoming popular, presumably on the basis of ease-of-computation and intuitive appeal.

As usual, take **deviance** to be $D(\theta) = -2 \log \pi(\text{data}|\theta)$.

Reflect **infidelity** of model to data by posterior mean deviance, $E\{D(\theta)|\text{data}\}$.

Reflect **complexity** of model, by *effective dimension*, $p_D = E\{D(\theta)|\text{data}\} - D\{E(\theta|\text{data})\}$.

Choose model minimizing **infidelity plus complexity**.

- black-box-ized (WinBUGS),
- 'effective' dimension very appropriate in many problems,
- still much active discussion of pros/cons,
- rapidly taking over the market - but for the right reasons?

## Some references

General discussion of principled Bayes model comparison:
Berger, Kass, Raftery, others

Separate-model approach to computation:
Various papers by Sid Chib (Washington U.)

Uber-model approach:
Reversible-jump MCMC - Green (1995, *Biometrika*)

DIC paper: Spiegelhalter *et. al.* (2002, *JRSS-B*).

## Module #4

**Further topics**

- Identification
- Performance of interval estimates

## Identification

In general a statistical model is nonidentified if multiple sets of parameter values correspond to the same distribution of observables.

Identification is often regarded as a minimal condition for a model to be 'sensible'.

On the other hand, realistic models for imperfect data are sometimes nonidentified, and simplifications to fix this are dubious.

E.g., take $(SN, SP)$ as known exactly rather than approximately.

E.g., assume $X_1^*, X_2^*$ conditionally independent given $X$.

## Choices may be:

1. Work with a nonidentified but realistic model
2. Work with an identified but unrealistic model
3. Give up

**Compare 1 and 2:**

- Which point estimator is more biased?
- Which interval estimator is more misleading?

## Working with a nonidentified model

No basis for computing or reporting ML inferences.

Bayesian inference: the math and computing involved in determining a posterior distribution is blind to whether or not the model is identified.

However, we expect the lack of identification to have an impact on the shape (particularly the concentration) of the posterior.

**Intuition???**

- posterior doesn't narrow to a single point as $n \to \infty$?
- posterior as wide as prior?

## Bayesian inference under nonidentified model

Have original model parameterization $\theta$ and prior $\pi$.

Sometimes, can find special parameterization $\phi = (\phi_I, \phi_N)$ isolating which terms do / don't appear in likelihood.

Often a sensible/weak prior in the original parameterization induces a strong dependence in the special parametrization.

Even support of $\phi_N$ depending on $\phi_I$ in some cases!

Let $g(\phi)$ be target parameter of interest.

Let $\tilde{g}(\phi_I) = E_\pi\{g(\phi)|\phi_I\}$.

Point estimation of target?

$E\{g(\phi)|\text{Data}\} = E\{\tilde{g}(\phi_I)|\text{data}\}$

where RHS is a 'regular' estimator of $\tilde{g}(\phi_I)$,
i.e., get consistent estimation of wrong target!

## Inference under nonidentified model, continued

Let $*$ denote true values. As $n \to \infty$:

$$\begin{aligned} E\{g(\phi)|\text{Data}\} &\to \tilde{g}(\phi_I^*) \neq g(\phi^*) \\ &\quad [\text{may be far from prior mode of } g(\phi)], \\ Var\{g(\phi)|\text{Data}\} &\to Var_\pi\{g(\phi)|\phi_I = \phi_I^*\} \\ &\quad [\leq Var_\pi\{g(\phi)\} \text{on average}]. \end{aligned}$$

Bayesian inference under a nonidentified model not (necessarily) useless.

Posterior variance not (necessarily) misleadingly narrow.

## Bias tradeoffs

**Smaller model:** identified but involves dubious assumptions.
**Bigger model:** more realistic but nonidentified.

Think of parameter governing departure from smaller model. As this moves away from 'zero' look at:

- bias due to nonidentification when using bigger model
- bias due to misspecification when using smaller model

Often the latter will quickly dominate!

## Example in brief

$$\begin{aligned} X^*|Y, X, S &\sim N(X, \tau^2) \\ Y|X, S &\sim N(\beta_0 + \beta_x X + \beta_s S, \sigma^2) \\ X|S &\sim N(\alpha_0 + \alpha_s S, \lambda^2) \end{aligned}$$

Want to infer $\beta_x$, but only get to observe $(X^*, Y, S)$.

Aided by $S$ being either exactly or approximately an instrumental variable.

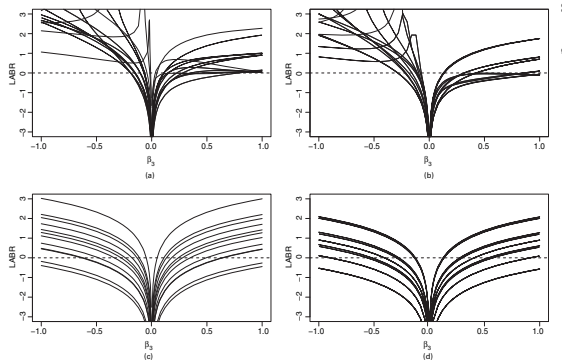
respsective priors

1. $\beta_s \equiv 0$: identified, risk of misspecified
2. $\beta_s \sim$ small: nonidentified, perhaps better specified.

## As the true value of $\beta_3$ moves away from zero...

How quickly does bias from misspecified model 1 become bigger than the bias using nonidentified model 2?

Log-absolute-bias-ratio:




## Interval coverage

**Valid frequentist confidence interval:**
For any fixed value of $\theta$, draw data from $\pi(\text{data}|\theta)$,
95% chance of generating a 'good' interval.

Typically requires a model that is both correct and identified.

**Valid Bayesian credible interval:**
Randomly draw $(\theta, \text{data})$ from $\pi(\theta)\pi(\text{data}|\theta)$,
95% chance of generating a 'good' interval.

A weaker notion of correct coverage,
can be achieved with a correct model and the 'right' prior.

## Interval coverage, continued

Think of nature's prior $\pi_{NAT}(\theta)$ and investigator's prior $\pi_{INV}(\theta)$.

**Look at coverage when data are:**

- generated from $\pi_{NAT}(\theta)\pi(\text{data}|\theta)$,
- analyzed using $\pi_{INV}(\theta)$, $\pi(\text{data}|\theta)$.

How fast does coverage deviate from nominal as $\pi_{NAT}$ deviates from $\pi_{INV}$?

Expect very minor deviations in the identified-model, large-sample setting.

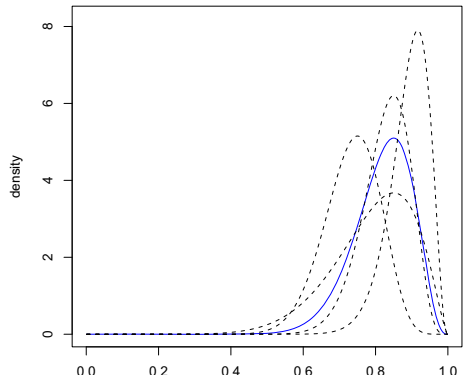
Concerned about very substantial deviations in nonidentified-model setting.

## Example

Case-control with exposure misclassification - again!

Consider different priors on $SN, SP$

$\pi_{NAT}$ is solid curve (mode at 0.85).



## Example, continued

Coverage of nominal 95% interval estimates for log-OR, under $\pi_{NAT}(\theta)\pi(\text{data}|\theta)$ data generation.

| | | |
|---|---|---|
| credible interval with $\pi_{INV} = \pi_{NAT}$ | 95% | |
| with $\pi_{INV} \neq \pi_{NAT}$ | 95% | |
| | 94% | |
| | 95% | |
| | 87% | |
| frequentist CI assuming $SN = SP = 1.00$: | 44% | |
| assuming $SN = SP = 0.85$: | 81% | (with caveat) |

## Some references

Various papers of Greenland and/or Gustafson, also Lawrence Joseph.

Neath and Samaniego (1997 *Am. Stat.*)
Poirier (1998, *Econometric. Th.*)
Xie and Carlin (2006, *JSPI*)

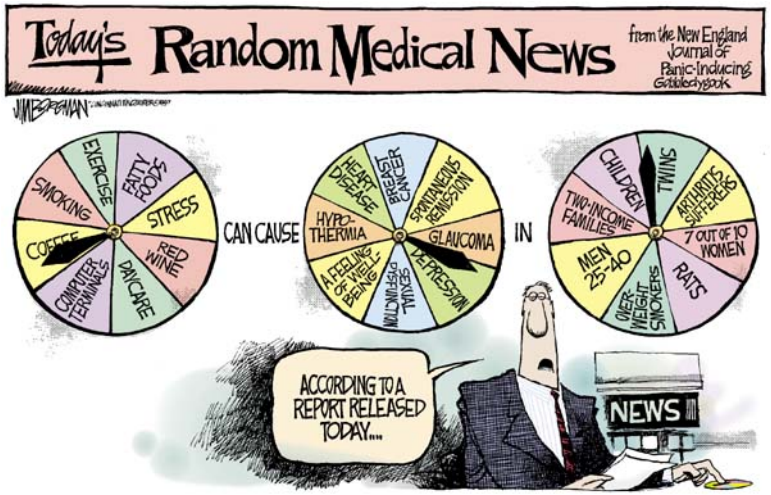
## Wrap-up

Bayes is delightfully simple (conceptually if not computationally) and intuitive.

Seems tailor-made for some problems.

Perhaps with more (Bayesian) acknowledgment of uncertainty we can progress away from the perceptions expressed in the following (old!) cartoon...

## Can Bayes save us from this???



Thanks for attending!