Feedback in Bayesian Models

Martyn Plummer

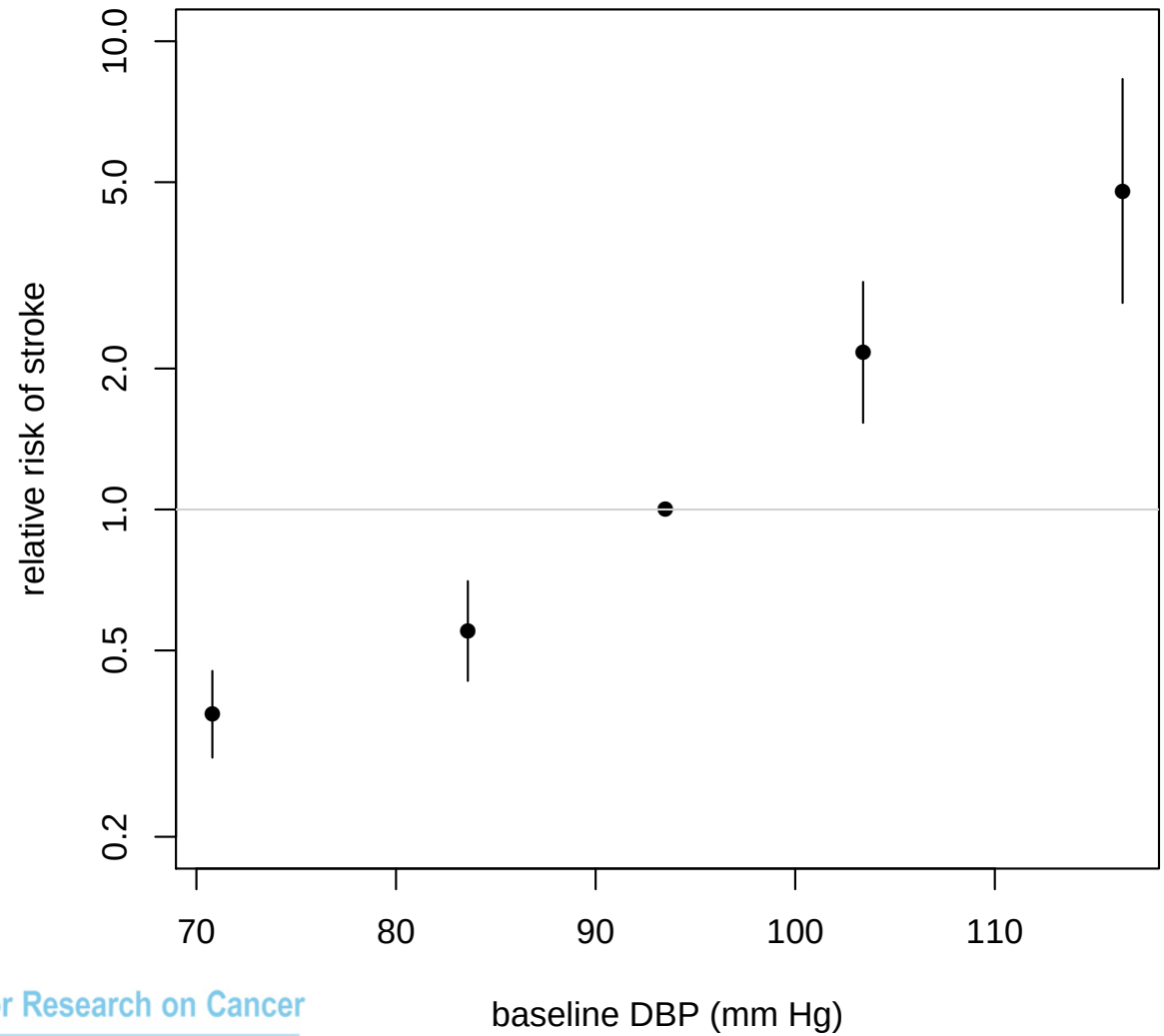
International Agency for Research on Cancer

Vienna Nov 20 2015

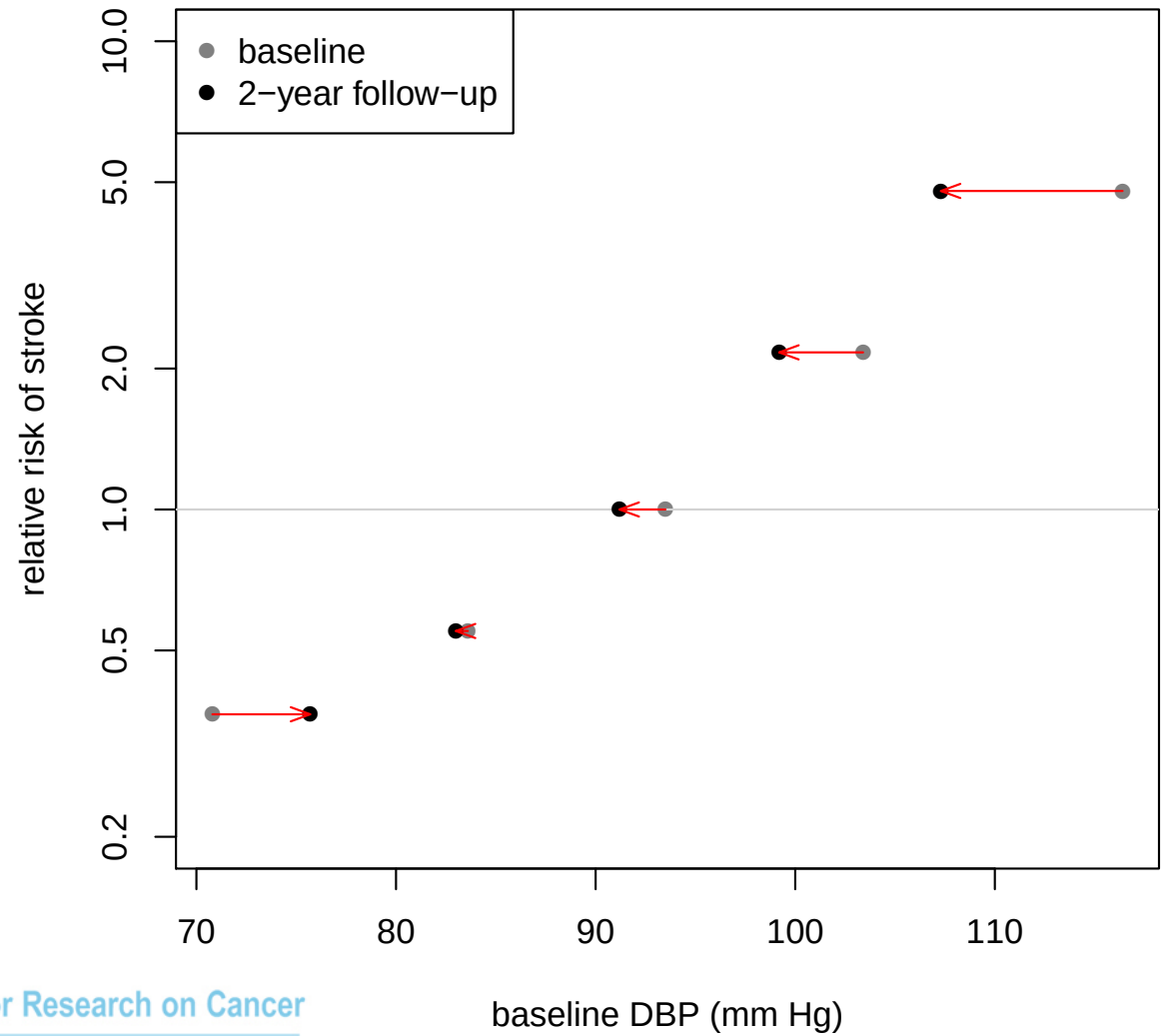
Motivation 1: measurement error in epidemiology

- MacMahon *et al* (1990): collaborative re-analysis of 9 cohort studies of blood pressure, stroke and coronary heart disease
- Participants categorized by baseline diastolic blood pressure (DBP) in 5 categories
  - ≤ 79; 80-89; 90-99; 100-108; ≥ 110 mm Hg
- What is the relationship between *average* DBP and stroke risk?

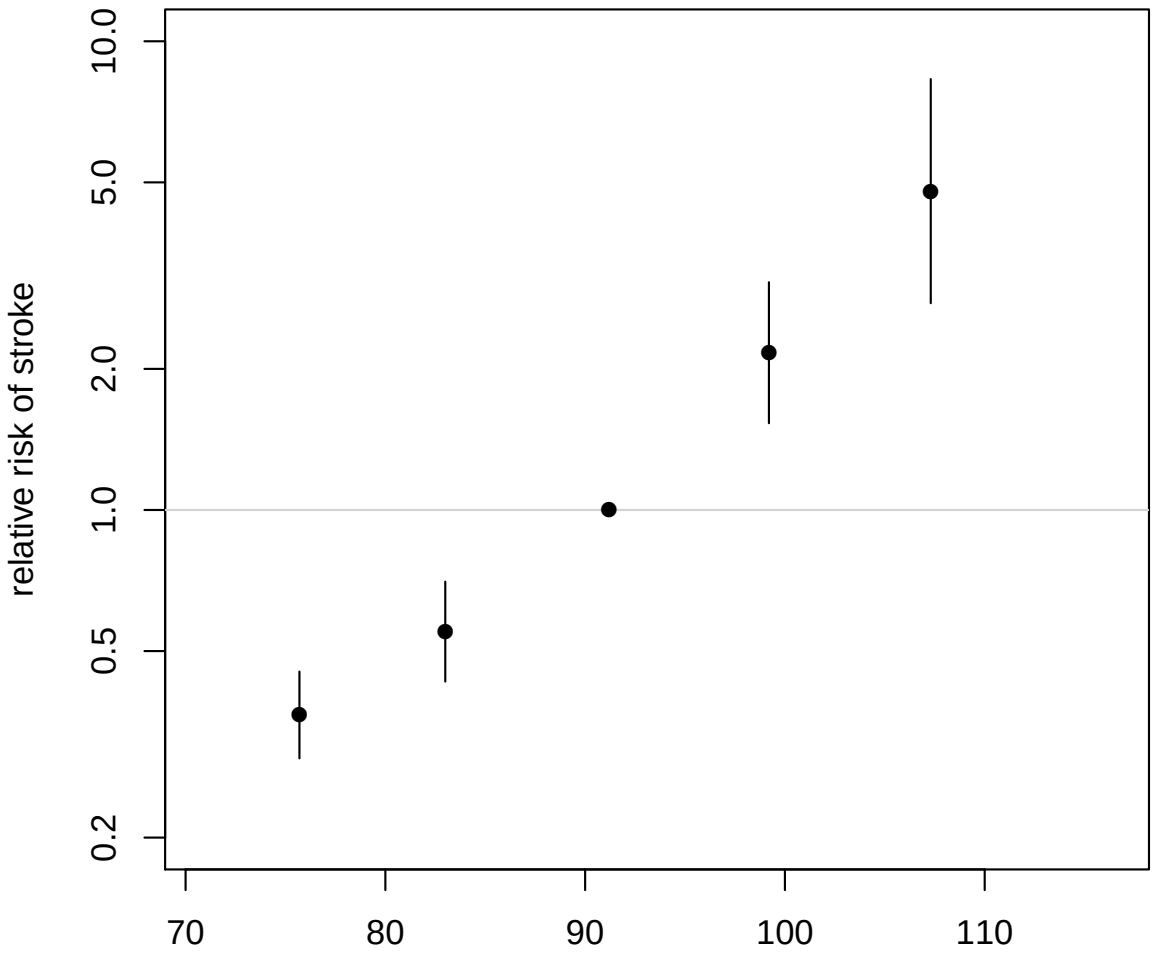
Stroke risk by baseline DBP



Regression to the mean in follow-up DBP measurement

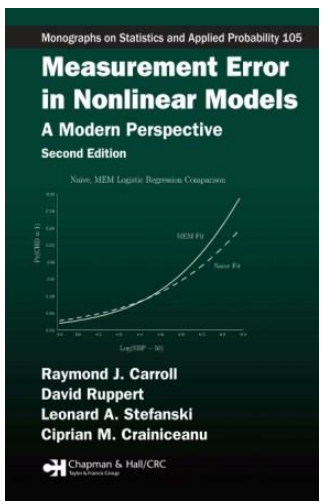


Stroke risk by mean DBP



Regression Calibration

Theory of regression calibration is now well-developed (See, for example, Carroll *et al* 2006)



- Calculate plug-in predictor values:
  - Expected value of true predictor given surrogate
  - Using data from validation/calibration sub-study
- Regress outcome on plug-in predictors
- Correct standard errors for uncertainty in plug-in predictors

Useful features of regression calibration

- Uses only data from calibration study for imputation of true predictors
- Does not depend on any assumptions about dose response
- In practice, efficient compared to full likelihood analysis

Is there a Bayesian equivalent to regression calibration?

Notation for linear model example

Calibration data: true exposure (X^*) and surrogate(Z^*)

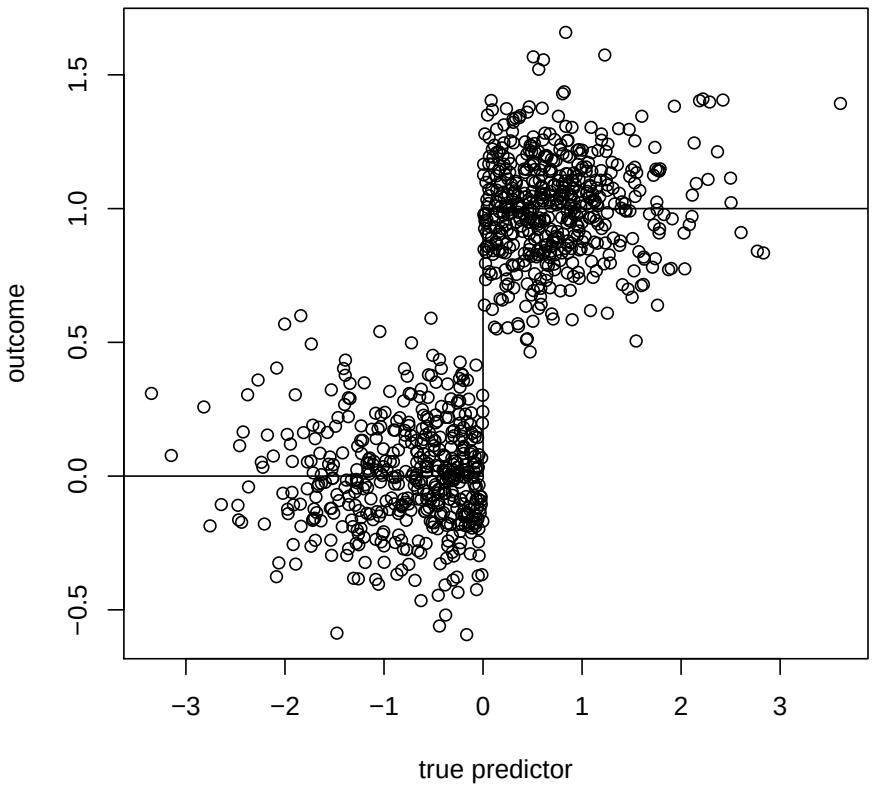
$$\begin{aligned} X_i^* &\sim \mathrm{N}(\mu_x, \tau_x^{-1}) && i = 1 \ldots m \\ Z_i^* \mid X_i^* &\sim \mathrm{N}(\alpha_z + \beta_z X_i^*, \tau_z^{-1}) && i = 1 \ldots m \end{aligned}$$

Regression data : surrogate(Z) and outcome (Y)

$$\begin{aligned} X_j &\sim \mathrm{N}(\mu_x, \tau_x^{-1}) && j = 1 \ldots N \\ Z_j \mid X_j &\sim \mathrm{N}(\alpha_z + \beta_z X_j, \tau_z^{-1}) && j = 1 \ldots N \\ Y_j \mid X_j &\sim \mathrm{N}(\alpha_y + \beta_y X_j, \tau_y^{-1}) && j = 1 \ldots N \end{aligned}$$

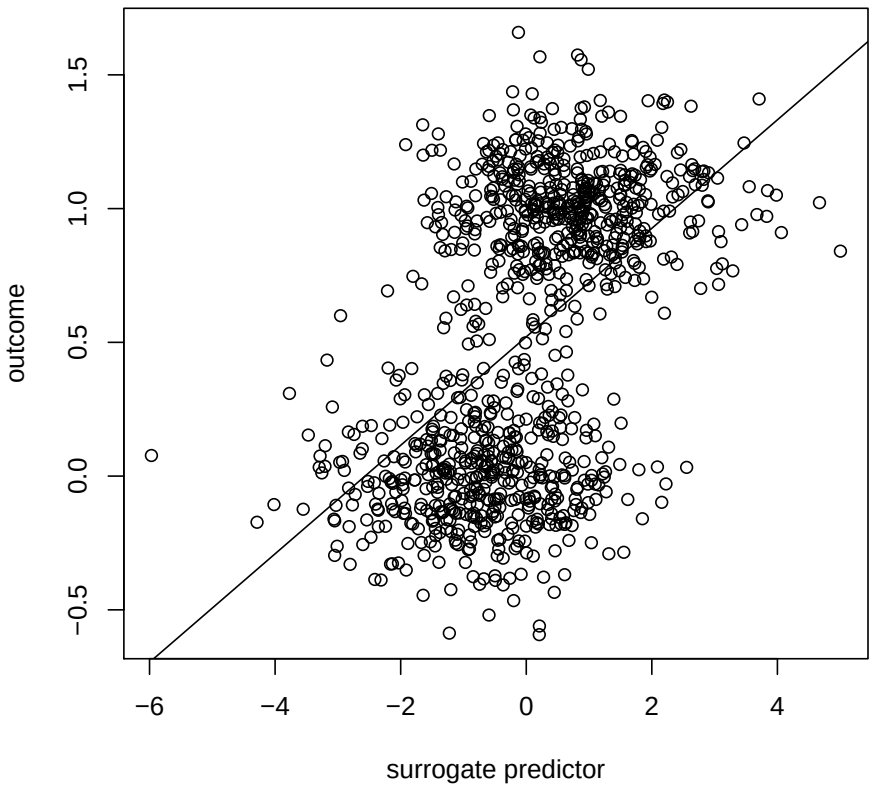
True exposure (X) is unobserved in regression data

Simulation: egregiously mis-specified dose-response



What happens if we fit a linear regression model to data generated by a threshold effect (or step-function)

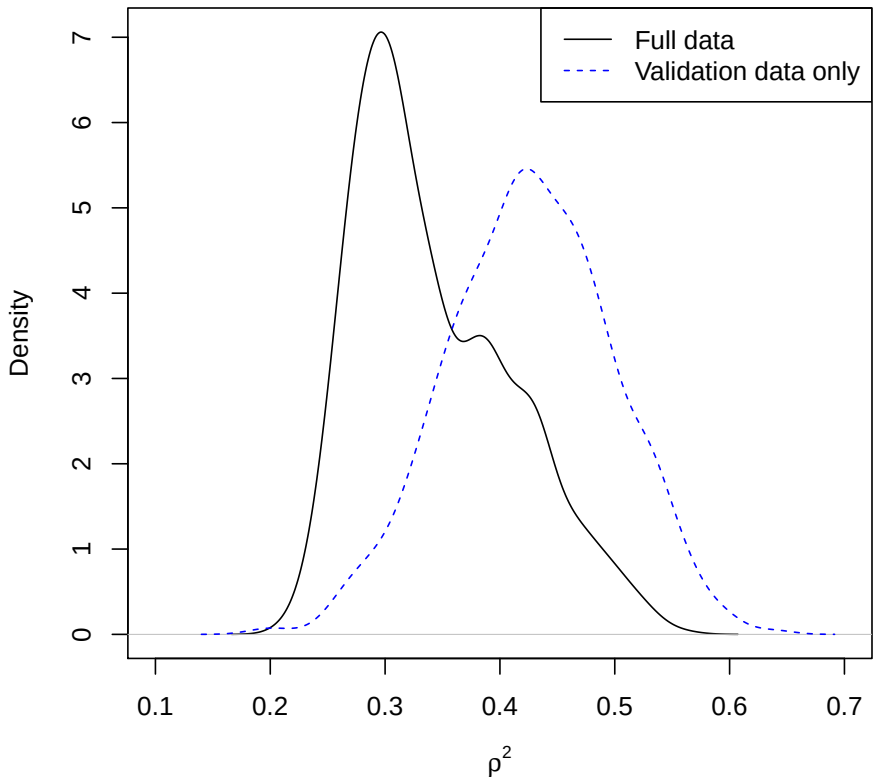
mis-specified dose-response by surrogate



The step-function is less obvious when using surrogate predictors. You could fit a linear regression but it diagnostic checks would show the mis-specification.

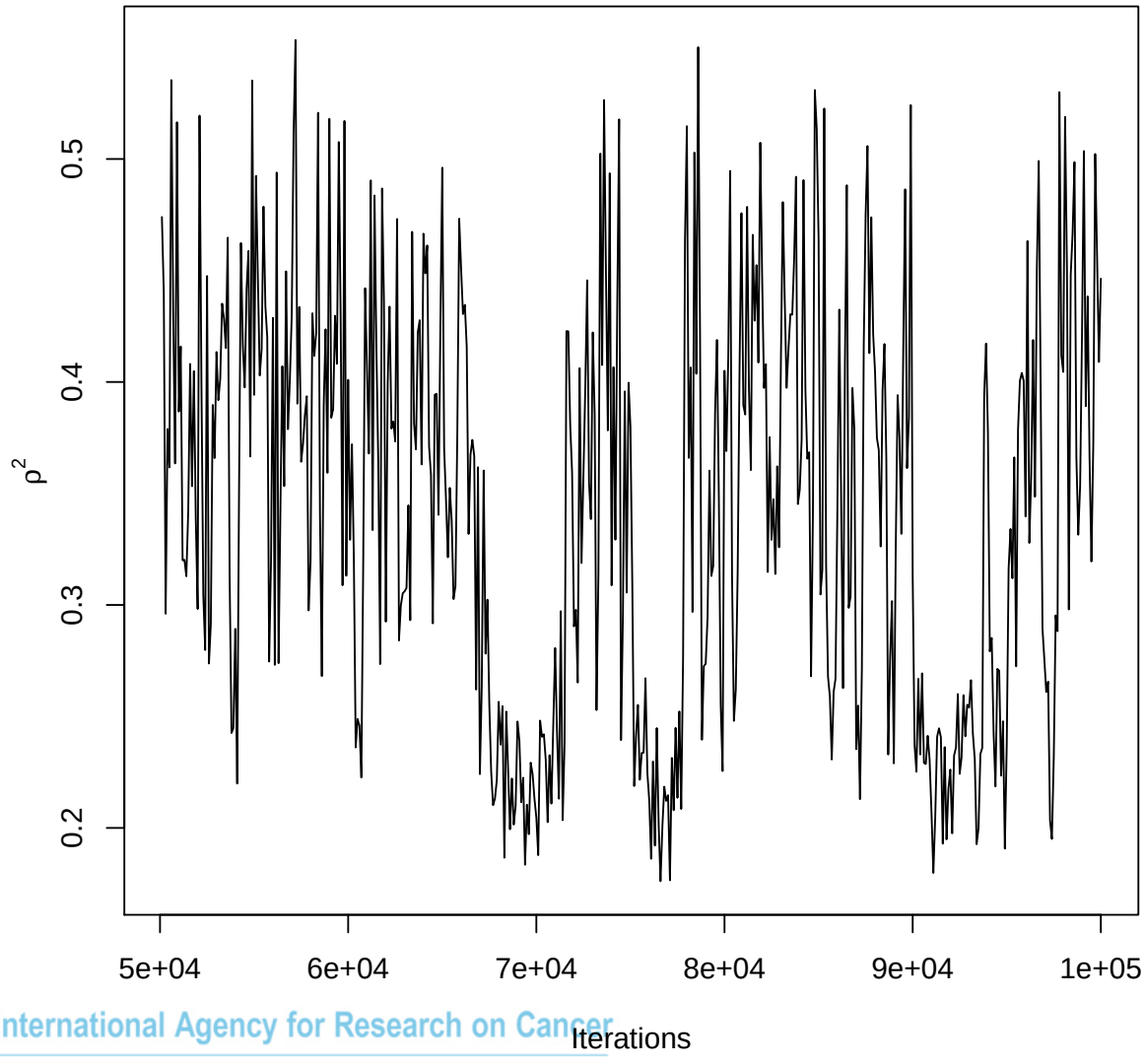
Feedback in a Bayesian full probability model

Quality of surrogate measurement determined by correlation between true and surrogate predictors (ρ).



Including outcome data and using the mis-specified linear regression model forces ρ to appear worse. Lunn *et al* (2009) call this phenomenon "feedback". Liu *et al* (2009) call it "contamination"

Feedback and MCMC mixing



Feedback is often accompanied by poor mixing of MCMC.
Here we have extremely high autocorrelation, and jumping between two local modes of the posterior for ρ^2.
Poor mixing is a strong motivation to seek alternate solutions.

Modularization

- A large model combining different data sources can be conceptually divided into "modules"
- Clayton (1992) described three sub-models of measurement error models in epidemiology:

  Exposure model Distribution of exposure in population

  Measurement model Relationship between true exposure and surrogate

  Disease model Relationship of disease outcome to true exposure
- Liu *et al* (2009) describe modified MCMC algorithms that weaken relationships between modules as "modularization".

Motivation 2: Population PK/PD

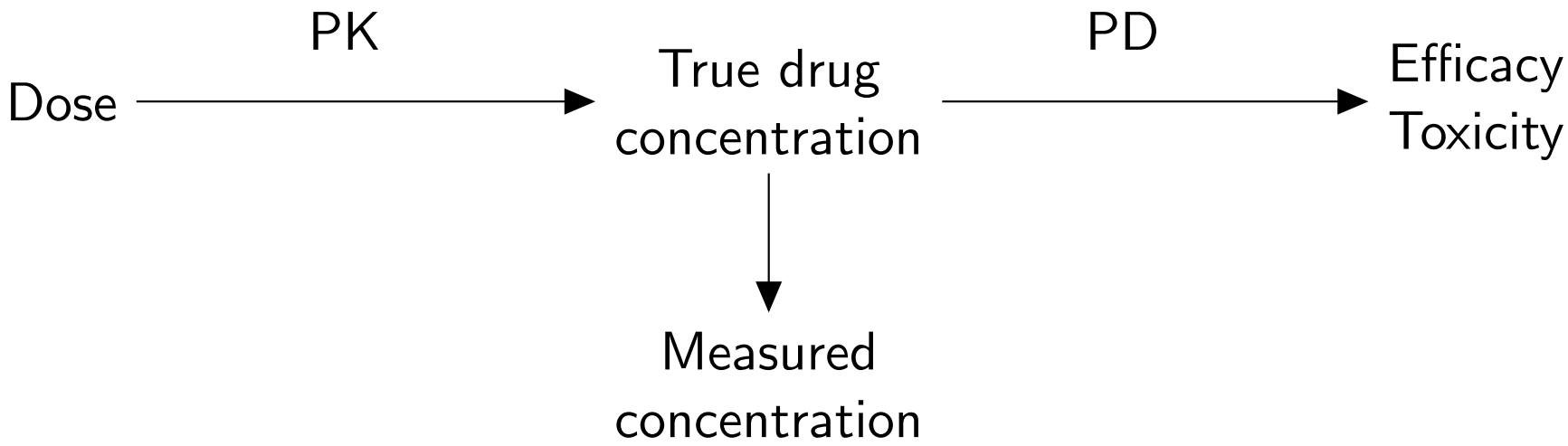
Population pharmacokinetic/pharmacodynamic (PD/PD) models aim to elicit the effects of drugs at a population level

- Variation within and between individuals
- Compartmental models
- Highly non-linear

NB Time dimension is missing in this graphical representation.

Measurement error in Population PK/PD

True concentration is not known exactly



Use PK model to get estimates of true drug concentration for PD model.

Sequential analysis of PK/PD models

Bennett & Wakefield (2001): Bayesian PK/PD model

- Insufficient PK data $\rightarrow$ under-fitted PK model
- Worse predictions than using observed concentration for PD model

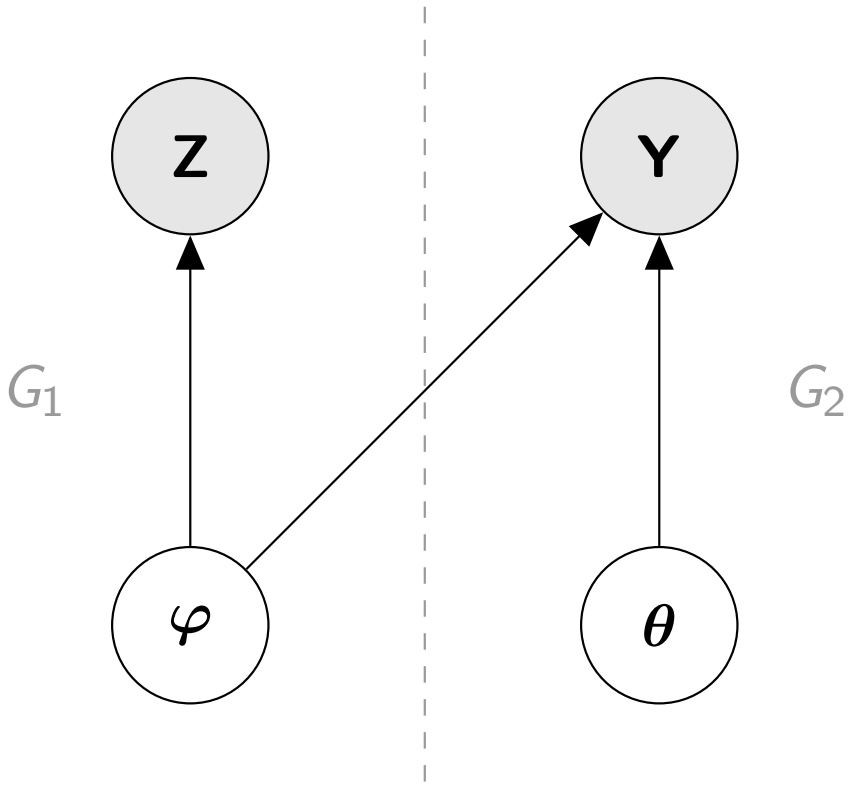
Zhang *et al* (2003a, 2003b): Frequentist sequential analysis

- Various strategies for plug-in concentration estimates based on PK only data
- Compared to simultaneous estimation: efficient; fast; robust to PD model mis-specification

Lunn *et al* (2009): Bayesian "sequential" PK/PD with MCMC

- Bayesian analogues of Zhang *et al* via "cuts"

A cut model



In a cut model, the graph G is divided into two sub-graphs G_1, G_2.

- Nodes in G_1 are updated ignoring nodes in G_2.
- Nodes in G_2 are updated as normal

Call this the *naive cut algorithm*

Implementation of cut models in OpenBUGS

OpenBUGS provides a cut function:

```
for (i in 1:N) {
   x.cut[i] <- cut(x[i])
}
```

When calculating full conditional distribution of `x[i]`, likelihood contributions from stochastic children of `x.cut[i]` are ignored.

> *The "cut" function forms a kind of 'valve' in the graph: prior information is allowed to flow 'downwards' through the cut, but likelihood information is prevented from flowing upwards.*
>
> *– OpenBUGS Manual*

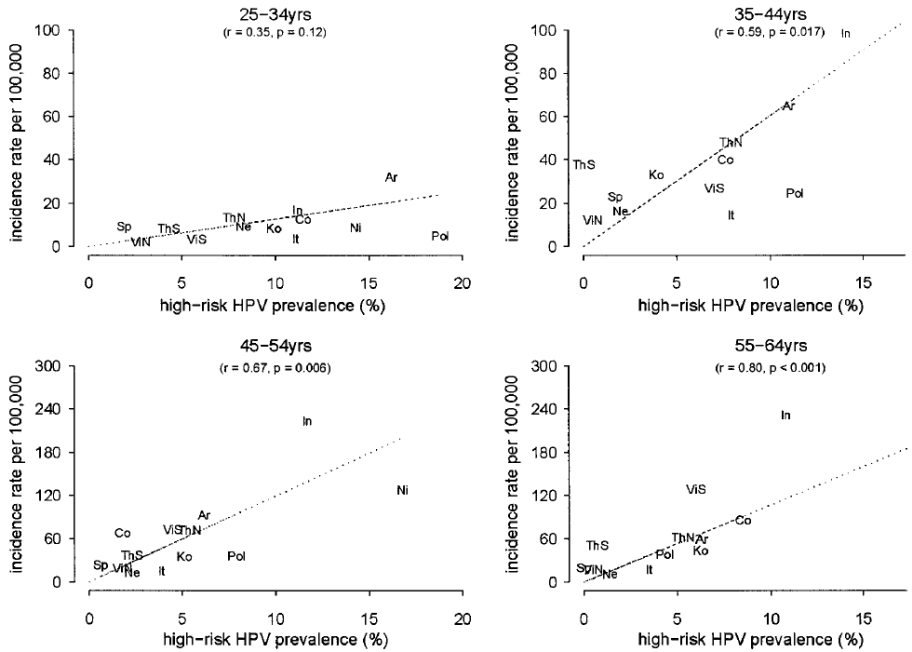
Represented using diode notation in a graph.

Other examples of modified MCMC algorithms

- Liu, Bayarri and Berger (2009) deal with contamination problem in computer models
- van Dyk and Jiao (2015) – "Partially Collapsed Gibbs Samplers"
  - Modify MCMC updates to ignore some information
  - But keep full posterior as target distribution
- Multiple Imputation with Chained Equations (MICE) for missing data.
  - Doubts often expressed about foundations when imputation models are incoherent

Toy epidemiological example

There is an ecological association between HPV prevalence and cervical cancer incidence[1]



HPV is a necessary cause of cancer, but risk is modulated by other cofactors: smoking, childbirth, hormonal contraceptives, ...

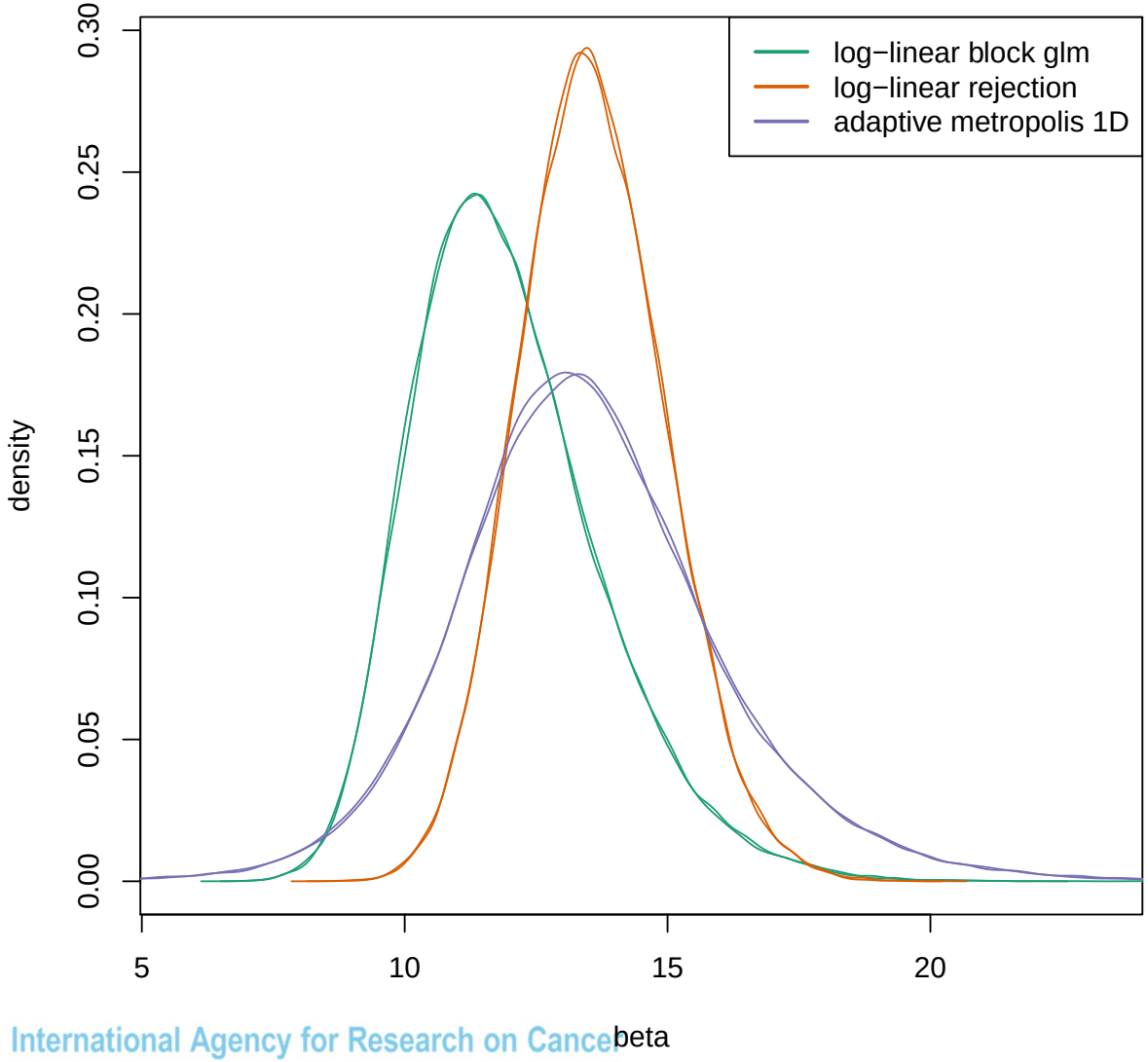
[1] Maucort-Boulch *et al* (2008)

A measurement error model for the ecological data

We experimented with a functional measurement error model for these data, with a Poisson regression model for incidence and a binomial model for (age-specific) prevalence:

$$
\begin{array}{lcll}
Y_i & \sim & \text{Poisson}(N_i \exp(\lambda_i)) & \text{Cancer incidence data} \\
\lambda_i & = & \theta_1 + \theta_2 \varphi_i & \text{Incidence rates} \\
Z_i & \sim & \text{Bin}(n_i, \varphi_i) & \text{HPV prevalence data}
\end{array}
$$

Results of naive cut algorithm for θ_2 by sampling method [2]



Different update methods converge to different limiting distributions.

[2] Plummer (2015), Statistics and Computing, **25**, 37–43

What is the target density of a cut model?

The target density of a cut model is the mixture:

$$p^*(\boldsymbol{\theta}) = \int p(\boldsymbol{\varphi} \mid \mathbf{Z}) p(\boldsymbol{\theta} \mid \boldsymbol{\varphi}, \mathbf{Y}) d\boldsymbol{\varphi}$$

This is the sampling density if we sample directly $\boldsymbol{\varphi}$ then $\boldsymbol{\theta}$ at each iteration.

Need to maintain this target distribution with other sampling schemes, e.g.

- Element-wise updating of $\boldsymbol{\varphi}, \boldsymbol{\theta}$
- Block-updating with reversible transitions

For convenience, consider block updating here.

Why the naive cut algorithm does not work

In general, MCMC methods do not sample directly from the target density but supply a reversible transition $\boldsymbol{\theta}^{t-1} \rightarrow \boldsymbol{\theta}^{t}$ at iteration t. The transition is in detailed balance with the full conditional distribution:

$$p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t}) p(\boldsymbol{\theta}^{t-1} \rightarrow \boldsymbol{\theta}^{t} \mid \boldsymbol{\varphi}^{t}) = p(\boldsymbol{\theta}^{t} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t}) p(\boldsymbol{\theta}^{t} \rightarrow \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^{t})$$

But for $p^{*}(\boldsymbol{\theta})$ to be the stationary distribution we need:

$$p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1}) p(\boldsymbol{\theta}^{t-1} \rightarrow \boldsymbol{\theta}^{t} \mid \boldsymbol{\varphi}^{t-1}, \boldsymbol{\varphi}^{t}) = p(\boldsymbol{\theta}^{t} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t}) p(\boldsymbol{\theta}^{t} \rightarrow \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^{t}, \boldsymbol{\varphi}^{t-1})$$

Why the naive cut algorithm does not work

In general, MCMC methods do not sample directly from the target density but supply a reversible transition $\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^{t}$ at iteration t. The transition is in detailed balance with the full conditional distribution:

$$p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t}) p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^{t} \mid \boldsymbol{\varphi}^{t}) = p(\boldsymbol{\theta}^{t} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t}) p(\boldsymbol{\theta}^{t} \to \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^{t})$$

But for $p^{*}(\boldsymbol{\theta})$ to be the stationary distribution we need:

$$p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1}) p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^{t} \mid \boldsymbol{\varphi}^{t-1}, \boldsymbol{\varphi}^{t}) = p(\boldsymbol{\theta}^{t} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t}) p(\boldsymbol{\theta}^{t} \to \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^{t}, \boldsymbol{\varphi}^{t-1})$$

The balance relation uses the current and previous values of $\boldsymbol{\varphi}$.

Can we modify a standard MCMC update? (1/2)

Maybe we can add a Metropolis-Hastings acceptance step, treating the move $\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^{t}$ as a proposal to be accepted with probability $\min(1, R)$ where

$$R = \frac{p(\boldsymbol{\theta}^{t} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t}) p(\boldsymbol{\theta}^{t} \to \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^{t-1})}{p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1}) p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^{t} \mid \boldsymbol{\varphi}^{t})}$$

Note that $R = 1$ in the case of direct sampling:

$$p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^{t} \mid \boldsymbol{\varphi}) = p(\boldsymbol{\theta}^{t} \mid \mathbf{Y}, \boldsymbol{\varphi})$$

Can we modify a standard MCMC update? (2/2)

For a standard MCMC update (in detailed balance with the full conditional distribution) the acceptance ratio can be rewritten in terms of forward transitions:

$$R = \frac{p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^t)}{p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1})} \frac{p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^{t-1})}{p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^t)}$$

But this requires

- Explicit expressions for the transition probabilities (not available for slice sampling, Hamiltonian Monte Monte Carlo).
- Evaluation of the ratio of two normalized densities

Can we modify a standard MCMC update? (2/2)

For a standard MCMC update (in detailed balance with the full conditional distribution) the acceptance ratio can be rewritten in terms of forward transitions:

$$R = \frac{p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^t)}{p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1})} \frac{p(\boldsymbol{\theta}^{t-1} \rightarrow \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^{t-1})}{p(\boldsymbol{\theta}^{t-1} \rightarrow \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^t)}$$

But this requires

- Explicit expressions for the transition probabilities (not available for slice sampling, Hamiltonian Monte Monte Carlo).
- Evaluation of the ratio of two normalized densities

Can we modify a standard MCMC update? (2/2)

For a standard MCMC update (in detailed balance with the full conditional distribution) the acceptance ratio can be rewritten in terms of forward transitions:

$$R = \frac{p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^t)}{p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1})} \frac{p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^{t-1})}{p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^t)}$$

But this requires

- Explicit expressions for the transition probabilities (not available for slice sampling, Hamiltonian Monte Monte Carlo).
- Evaluation of the ratio of two normalized densities
  - Unsuitable for most applications of MCMC where we have only unnormalized densities.

Numerical issues

- Use of cuts often motivated by numerical issues
  - Liu *et al* (2009) – not sufficient reason to modularize inference
- van Dyk and Jiao (2015) – sensitivity of Partially Collapsed Gibbs Samplers to update order and use of Metropolis-Hastings
- MICE?

Statistical issues

- Cuts represent a *refusal to learn* about certain parameters in the model
  - Lunn *et al* (2009) call these "distributional constants"
- Even if multiple imputation is a target for cut models, it leads to inconsistent inference
  - Meng (1994) Multiple imputation inferences with uncongenial sources of input
  - Nielsen (2003) Proper and improper multiple imputation

"Sequential" Bayesian analysis

- In practice "sequential" Bayesian analysis is used whenever we include prior distributions based on *summary statistics* from previous studies.
- Perhaps the "feedback" problem is due to trying to carry over the full posterior from stage 1 (calibration data only) to stage 2 (including surrogate and outcome regression data) instead of a simplified summary.

"Sequential Bayesian analysis

- In our Bayesian measurement error model, the predictive distribution of true predictor variables $X_1 \ldots X_N$ from stage 1 becomes the prior of $X_1 \ldots X_N$ in stage 2
- Hypothesis: We can reduce feedback by
  1. Keeping correct marginal predictive distribution of X_i from stage 1
  2. But treating X_i and X_j as *a priori* independent for $i \neq j$ in stage 2

  and *otherwise respecting normal rules for Bayesian inference and MCMC*

Stage 1

Replicate calibration data N times,

$$\begin{aligned} X^*_{ij} &\sim \mathrm{N}(\mu_{xj}, \tau^{-1}_{xj}) && i = 1 \ldots m \quad j = 1 \ldots N \\ Z^*_{ij} \mid X^*_{ij} &\sim \mathrm{N}(\alpha_{zj} + \beta_{zj} X^*_{ij}, \tau^{-1}_{zj}) && i = 1 \ldots m \quad j = 1 \ldots N \end{aligned}$$

Each copy has its own private parameters for

1. exposure model: μ_{xj}, τ_{xj}
2. measurement model: $\alpha_{zj}, \beta_{zj}, \tau_{zj}$

Hence, *e.g.* α_{zj} is independent of α_{zk} for $j \neq k$, also *a posteriori*.

Stage 2

Each observation in the regression data uses its own copy of the parameters from stage 1.

$$\begin{aligned} X_j &\sim \mathrm{N}(\mu_{xj}, \tau_{xj}^{-1}) && j = 1 \ldots N \\ Z_j \mid X_j &\sim \mathrm{N}(\alpha_{zj} + \beta_{zj} X_j, \tau_{zj}^{-1}) && j = 1 \ldots N \end{aligned}$$

Regression parameters are common

$$Y_j \mid X_j \sim \mathrm{N}(\alpha_y + \beta_y X_j, \tau_y^{-1}) \quad j = 1 \ldots N$$

How does this affect feedback?

- Marginal posterior of X_i given only validation data $\mathbf{X}^*, \mathbf{Z}^*$ is the same as in full probability model
- Parameters of exposure model and measurement model are estimated from m validation measurements but only 1 outcome measurement
  - Informally, influence of outcome data on distribution of X_i is $O(m^{-1})$ not $O(Nm^{-1})$
- Data replication is computationally expensive, but there is scope for parallelization

What do I hope to see?

Some kind of efficiency/robustness trade-off, *e.g.*

- Minimal loss of efficiency when model is true (*q.v.* regression calibration)
- Robustness to outliers
- Increased ability to detect mis-specified outcome model by posterior predictive simulation

Perspectives

- Cuts take an *algorithmic* view of the feedback problem
- Statistical properties not well defined
- Promoted by software implementation

References

- J Bennett, J Wakefield (2001) Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling, *Biometrics*, **57**, 803–812
- RJ Carroll, D Ruppert, LA Stefanski, C Crainiceanu (2006) *Measurement error in nonlinear models, second edition*, Chapman & Hall/CRC
- DG Clayton (1992) Models for the longitudinal analysis of cohort and case-control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health* (eds JH Dwyer, M Feinleib, P Lippert, H Hoffmeister), 301–331. Oxford: Oxford University Press
- F Liu, MJ Bayarri, JO Berger (2009) Modularization in Bayesian analysis with emphasis on analysis of computer models, *Bayesian Analysis*, **4**, 119–150
- D Lunn, N Best, D Spiegelhalter, G Graham, B Neuenschwander (2009) Combining MCMC with 'sequential' PKPD modelling, *J Pharmacokinet Pharmacodynam*, **36**, 19–38
- S MacMahon, R Peto, J Cutler, R Collins, P Sorlie, J Neaton, R Abbott, J Godwin, A Dyer, J Stamler (1990) Blood pressure, stroke, and coronary heart disease: Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias, *Lancet*, **355**, 765–74
- D Maucort-Boulch, S Franceschi, M Plummer; IARC HPV Prevalence Surveys Study Group (2008) International correlation between human papillomavirus prevalence and cervical cancer incidence, *Cancer Epidemiol Biomarkers Prev*, **17**, 717–20
- X-L Meng (1994) Multiple imputation inference with uncongenial sources of input, *Statistical Science*, **9**, 538–573
- SF Nielsen (2003) Proper and improper multiple imputation, *International Statistical Review*, **71**, 593–627
- M Plummer (2015) Cuts in Bayesian graphical models, *Stat Comput*, **25**, 37–43
- D van Dyk, X Jiao (2015) Metropolis-Hastings with partially collapsed Gibbs samplers, *J Comput Graph Stat*, **24**, 301–327
- L Zhang, SL Beal, LB Sheiner (2003a) Simultaneous vs sequential analysis for population PK/PD data I: Best-case performance, J Pharmacokinet Pharmacodynam, **30**, 387–404
- L Zhang, SL Beal, LB Sheiner (2003b) Simultaneous vs sequential analysis for population PK/PD data II: Robustness of methods, J Pharmacokinet Pharmacodynam, **30**, 405–416