Bayesian inference with models made of modules

Pierre E. Jacob

EcoDep conference
June 23, 2022

Outline

Outline

Models made of modules

- First module:

  parameter θ_1, data Y_1

  prior: $p_1(\theta_1)$

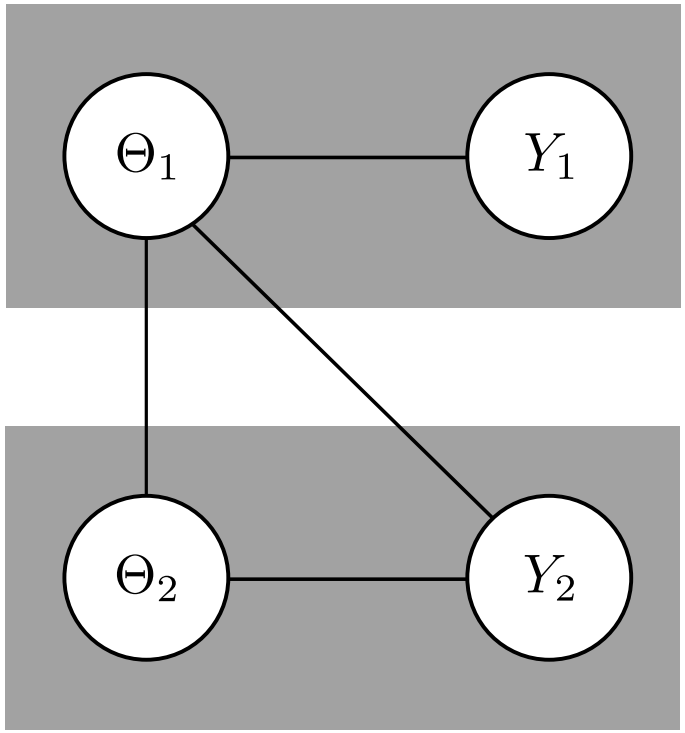
  likelihood: $p_1(Y_1|\theta_1)$

- Second module:

  parameter θ_2, data Y_2

  prior: $p_2(\theta_2|\theta_1)$

  likelihood: $p_2(Y_2|\theta_1, \theta_2)$



Joint model approach

Parameter (θ_1, θ_2), with prior

$$p(\theta_1, \theta_2) = p_1(\theta_1) p_2(\theta_2|\theta_1).$$

Data (Y_1, Y_2), likelihood

$$p(Y_1, Y_2|\theta_1, \theta_2) = p_1(Y_1|\theta_1) p_2(Y_2|\theta_1, \theta_2).$$

Posterior distribution

$$\pi(\theta_1, \theta_2|Y_1, Y_2) \propto p_1(\theta_1)\, p_1(Y_1|\theta_1) p_2(\theta_2|\theta_1)\, p_2(Y_2|\theta_1, \theta_2).$$

Joint model approach

Parameter (θ_1, θ_2), with prior

$$p(\theta_1, \theta_2) = p_1(\theta_1) p_2(\theta_2|\theta_1).$$

Data (Y_1, Y_2), likelihood

$$p(Y_1, Y_2|\theta_1, \theta_2) = p_1(Y_1|\theta_1) p_2(Y_2|\theta_1, \theta_2).$$

Posterior distribution

$$\pi(\theta_1, \theta_2|Y_1, Y_2) \propto p_1(\theta_1)\, p_1(Y_1|\theta_1) p_2(\theta_2|\theta_1)\, p_2(Y_2|\theta_1, \theta_2).$$

Departures from the joint model approach can be sensible.

Example: biased data

Liu, Bayarri & Berger (2009). *Modularization in Bayesian analysis, with emphasis on analysis of computer models.*

- Location model:

  $\forall i = 1, \ldots, n_1 \quad Y_1^i \sim \text{Normal}(\theta_1, 1)$

  $\theta_1 \sim \text{Normal}(0, 1)$

- Extra data Y_2 suspected to be biased:

  $\forall i = 1, \ldots, n_2 \quad Y_2^i \sim \text{Normal}(\theta_1 + \theta_2, 1)$

  $\theta_2 \sim \text{Normal}(0, v)$

Example: biased data

Liu, Bayarri & Berger (2009). *Modularization in Bayesian analysis, with emphasis on analysis of computer models.*

- Location model:

  $\forall i = 1, \ldots, n_1 \quad Y_1^i \sim \text{Normal}(\theta_1, 1)$

  $\theta_1 \sim \text{Normal}(0, 1)$

- Extra data Y_2 suspected to be biased:

  $\forall i = 1, \ldots, n_2 \quad Y_2^i \sim \text{Normal}(\theta_1 + \theta_2, 1)$

  $\theta_2 \sim \text{Normal}(0, v)$

If interest is in θ_1: are the extra data useful or harmful?

If interest is in θ_2: joint model or "two-step" approach?

Example: COVID-19 prevalence

Nicholson et al. (2022). *Interoperability of statistical models in pandemic preparedness: principles and reality.*

Prevalence π of SARS-COV-2 in the UK, estimated from

- randomized surveillance data: u positive out of U tested, Hypergeometric model with parameter π.
- targeted surveillance data (patients with clinical need, health & care workers): n positive out of N tested, Binomial model involving π, $\mathbb{P}(\text{tested}|\text{infected})$ and $\mathbb{P}(\text{tested}|\text{not infected})$.

Note: $U \ll N$.

Example: COVID-19 prevalence

Nicholson et al. (2022). *Interoperability of statistical models in pandemic preparedness: principles and reality.*

Prevalence π of SARS-COV-2 in the UK, estimated from

- randomized surveillance data: u positive out of U tested, Hypergeometric model with parameter π.
- targeted surveillance data (patients with clinical need, health & care workers): n positive out of N tested, Binomial model involving π, $\mathbb{P}(\text{tested}|\text{infected})$ and $\mathbb{P}(\text{tested}|\text{not infected})$.

Note: $U \ll N$.

If interest is in π: are the extra data useful or harmful?

If interest is in e.g. $\mathbb{P}(\text{tested}|\text{infected})$: joint model or "two-step" approach?

Example: PKPD

Lunn, Best, Spiegelhalter, Graham & Neuenschwander (2009).
Combining MCMC with 'sequential' PKPD modelling.

- Pharmacokinetics (PK):
  models the time course of drug absorption.

  $\forall t \quad Y_t \sim \text{Normal}(\log C_t, v_{\text{PK}}), \text{ where } C_t = \text{function}(t, \theta_{\text{PK}}).$

  From this we extract $(C_t^{(j)})_{t \geq 0}$ for individual $j = 1, \ldots, J$.
- Pharmacodynamics (PD):
  models the effect of drugs.

  $\forall j \quad Z_j \sim \text{Normal}(E_j, v_{\text{PD}}), \text{ where } E_j = \text{function}(C_{t_j}^{(j)}, \theta_{\text{PD}}),$

  and where t_j is the time at which E_j is measured.

Example: PKPD

Lunn, Best, Spiegelhalter, Graham & Neuenschwander (2009).
Combining MCMC with 'sequential' PKPD modelling.

- Pharmacokinetics (PK):
  models the time course of drug absorption.

  $\forall t \quad Y_t \sim \text{Normal}(\log C_t, v_{\text{PK}}), \text{ where } C_t = \text{function}(t, \theta_{\text{PK}}).$

  From this we extract $(C_t^{(j)})_{t \geq 0}$ for individual $j = 1, \ldots, J$.
- Pharmacodynamics (PD):
  models the effect of drugs.

  $\forall j \quad Z_j \sim \text{Normal}(E_j, v_{\text{PD}}), \text{ where } E_j = \text{function}(C_{t_j}^{(j)}, \theta_{\text{PD}}),$

  and where t_j is the time at which E_j is measured.

Interest in θ_{PD} ultimately, but more trust in PK model.

Example: epidemiological study

Plummer (2014). *Cuts in Bayesian graphical models.*

- Human papillomavirus prevalence φ_i in country i:

$$\forall i = 1, \ldots, I \quad Z_i \sim \text{Binomial}(N_i, \varphi_i),$$

Z_i: number of women infected with high-risk HPV,

N_i: population size in country i.

- Impact of prevalence onto cervical cancer occurrence:

$$\forall i = 1, \ldots, I \quad Y_i \sim \text{Poisson}(\lambda_i T_i), \quad \log(\lambda_i) = \theta_{2,1} + \theta_{2,2}\varphi_i,$$

Y_i is number of cases during study in country i,

T_i: woman-years of follow-up in country i.

Example: epidemiological study

Plummer (2014). *Cuts in Bayesian graphical models.*

- Human papillomavirus prevalence φ_i in country i:

$$\forall i = 1, \ldots, I \quad Z_i \sim \text{Binomial}(N_i, \varphi_i),$$

Z_i: number of women infected with high-risk HPV,

N_i: population size in country i.

- Impact of prevalence onto cervical cancer occurrence:

$$\forall i = 1, \ldots, I \quad Y_i \sim \text{Poisson}(\lambda_i T_i), \quad \log(\lambda_i) = \theta_{2,1} + \theta_{2,2}\varphi_i,$$

Y_i is number of cases during study in country i,

T_i: woman-years of follow-up in country i.

Interest in $\theta_{2,2}$, easier to interpret once $(\varphi_i)_{i=1}^I$ are fixed.

Example: two-step regressions

Murphy & Topel (1985). *Estimation and Inference in Two-Step Econometric Models.*

Impact of unanticipated money growth on unemployment.

- $$\forall t \quad \mathrm{DM}_t = X_{1t}\theta + \mathrm{DMR}_t,$$

  DM_t: proportional growth in the M1 definition of money,

  X_{1t}: lagged DM_t, lagged unemployment, more variables.

- $$\forall t \quad \log \frac{\mathrm{UN}_t}{1 - \mathrm{UN}_t} = X_{2t}\beta + \gamma(L)\mathrm{DMR}_t + u_t,$$

  UN_t: annual average unemployment rate,

  X_{2t}: minimum wage, more variables,

  $\gamma(L)$: 2nd order polynomial of lag operator.

Example: two-step regressions

Murphy & Topel (1985). *Estimation and Inference in Two-Step Econometric Models.*

Impact of unanticipated money growth on unemployment.

- $$\forall t \quad \mathrm{DM}_t = X_{1t}\theta + \mathrm{DMR}_t,$$

  DM_t: proportional growth in the M1 definition of money,

  X_{1t}: lagged DM_t, lagged unemployment, more variables.

- $$\forall t \quad \log \frac{\mathrm{UN}_t}{1 - \mathrm{UN}_t} = X_{2t}\beta + \gamma(L)\mathrm{DMR}_t + u_t,$$

  UN_t: annual average unemployment rate,

  X_{2t}: minimum wage, more variables,

  $\gamma(L)$: 2nd order polynomial of lag operator.

Interest in $\gamma(L)$, but requires imputation of unobserved DMR_t.

Joint estimation "inappropriate or computationally infeasible".

Example: state space models

Parslow, Cressie, Campbell, Jones & Murray (2013). *Bayesian learning and predictability in a stochastic nonlinear dynamical model.*

- Geophysics model of the temperature of the ocean, ϕ.
- The temperature ϕ can be used as "forcings" in a model of plankton population size β, for example in an SDE model

$$d\beta_t = \mu(\beta_t, \phi_t)dt + \sigma(\beta_t, \phi_t)dW_t.$$

Example: state space models

Parslow, Cressie, Campbell, Jones & Murray (2013). *Bayesian learning and predictability in a stochastic nonlinear dynamical model.*

- Geophysics model of the temperature of the ocean, ϕ.
- The temperature ϕ can be used as "forcings" in a model of plankton population size β, for example in an SDE model

$$d\beta_t = \mu(\beta_t, \phi_t)dt + \sigma(\beta_t, \phi_t)dW_t.$$

We might want to:

- propagate uncertainty about the geophysics to the biology?
- allow/prevent feedback from the biology to the geophysics?

Examples. . .

- Environmental epidemiology: estimation of environmental exposure, then associated health effects.
- Missing data: imputation of missing values, then analysis of completed data.
- Causal inference with propensity scores: estimation of probability of individuals receiving treatment, then treatment effect adjusted for propensity score.

Examples. . .

- Environmental epidemiology: estimation of environmental exposure, then associated health effects.
- Missing data: imputation of missing values, then analysis of completed data.
- Causal inference with propensity scores: estimation of probability of individuals receiving treatment, then treatment effect adjusted for propensity score.

And many more!

Jacob, Murray, Holmes & Robert (2017). *Better together? Statistical learning in models made of modules.*

Supporters say aye, opponents say no

Setup: model 2 depends on an input that is itself estimated using model 1.

Bayesian analysis with the joint model:

- coherency, simultaneous treatment of uncertainty, and other appeals of standard Bayes,
- computational toolbox is already available,

Supporters say aye, opponents say no

Setup: model 2 depends on an input that is itself estimated using model 1.

Bayesian analysis with the joint model:

- ☺ coherency, simultaneous treatment of uncertainty, and other appeals of standard Bayes,
- ☺ computational toolbox is already available,

- ☹ computationally challenging

  as difficulties pile up with more modules,
- ☹ parameters might be hard to interpret as their meaning changes across modules,
- ☹ module misspecification means that incorporating more data is not necessarily beneficial, and sometimes harmful.

Outline

Plug-in approach

Simple:

1. estimate θ_1 given Y_1 first, e.g. $\hat{\theta}_1 = \int \theta_1 \, p_1(\theta_1 | Y_1) d\theta_1$,
2. inference on θ_2 given Y_2 and $\hat{\theta}_1$ using

$$p_2(\theta_2 | \hat{\theta}_1, Y_2) = \frac{p_2(\theta_2 | \hat{\theta}_1) p_2(Y_2 | \hat{\theta}_1, \theta_2)}{p_2(Y_2 | \hat{\theta}_1)}.$$

Plug-in approach

Simple:

1. estimate θ_1 given Y_1 first, e.g. $\hat{\theta}_1 = \int \theta_1 \, p_1(\theta_1 | Y_1) d\theta_1$,
2. inference on θ_2 given Y_2 and $\hat{\theta}_1$ using

$$p_2(\theta_2 | \hat{\theta}_1, Y_2) = \frac{p_2(\theta_2 | \hat{\theta}_1) p_2(Y_2 | \hat{\theta}_1, \theta_2)}{p_2(Y_2 | \hat{\theta}_1)}.$$

☹ Uncertainty about θ_1 is ignored in the estimation of θ_2.

Cut approach

One might want to propagate uncertainty without allowing "feedback" of second module on first module.

Define the cut distribution:

$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) = p_1(\theta_1|Y_1) p_2(\theta_2|\theta_1, Y_2).$$

Bayesian version of two-step estimation.

Cut approach

One might want to propagate uncertainty without allowing "feedback" of second module on first module.

Define the cut distribution:

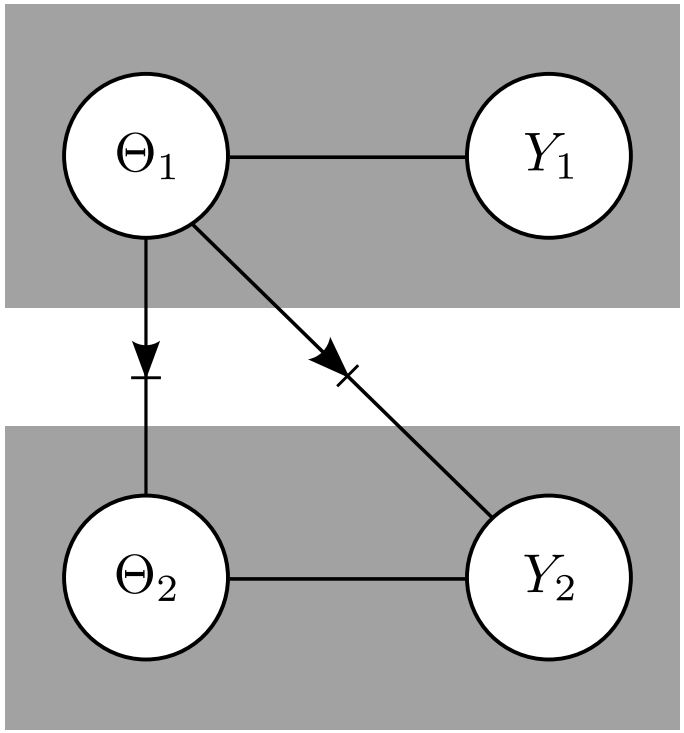
$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) = p_1(\theta_1|Y_1) p_2(\theta_2|\theta_1, Y_2).$$

Bayesian version of two-step estimation.

Ideal sampling procedure:

1. Sample θ_1 from $p_1(\theta_1|Y_1)$.
2. Given θ_1, sample θ_2 from $p_2(\theta_2|\theta_1, Y_2)$.
3. Output (θ_1, θ_2).

Cut approach



From the OpenBUGS manual,
Spiegelhalter, Thomas, Best & Lunn, 2004:

> *The cut function acts as 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.*

Cut approach

Difference between cut and standard posterior:

$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) \propto p_1(\theta_1) p_1(Y_1|\theta_1) \frac{p_2(\theta_2|\theta_1) p_2(Y_2|\theta_1, \theta_2)}{p_2(Y_2|\theta_1)}$$
$$\propto \frac{\pi(\theta_1, \theta_2|Y_1, Y_2)}{p_2(Y_2|\theta_1)}.$$

Cut approach

Difference between cut and standard posterior:

$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) \propto p_1(\theta_1) p_1(Y_1|\theta_1) \frac{p_2(\theta_2|\theta_1) p_2(Y_2|\theta_1, \theta_2)}{p_2(Y_2|\theta_1)}$$
$$\propto \frac{\pi(\theta_1, \theta_2|Y_1, Y_2)}{p_2(Y_2|\theta_1)}.$$

The term $p_2(Y_2|\theta_1)$ is a measure of feedback of Y_2 onto θ_1:

$$p_2(Y_2|\theta_1) = \int p_2(Y_2|\theta_1, \theta_2) p_2(\theta_2|\theta_1) d\theta_2.$$

Cut approach

Difference between cut and standard posterior:

$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) \propto p_1(\theta_1) p_1(Y_1|\theta_1) \frac{p_2(\theta_2|\theta_1) p_2(Y_2|\theta_1, \theta_2)}{p_2(Y_2|\theta_1)}$$
$$\propto \frac{\pi(\theta_1, \theta_2|Y_1, Y_2)}{p_2(Y_2|\theta_1)}.$$

The term $p_2(Y_2|\theta_1)$ is a measure of feedback of Y_2 onto θ_1:

$$p_2(Y_2|\theta_1) = \int p_2(Y_2|\theta_1, \theta_2) p_2(\theta_2|\theta_1) d\theta_2.$$

Another view on comparing cut and standard posterior:

- the marginal distribution of θ_1 differs,

  $p_1(\theta_1|Y_1)$ for cut, $\pi(\theta_1|Y_1, Y_2)$ for standard posterior,

- the conditional distribution of θ_2 is the same: $\pi(\theta_2|\theta_1, Y_2)$.

Variants

- Instead of completely cutting feedback, *semi-modular inference* controls the feedback, with a parameter in $[0, 1]$ to interpolate between the cut and the full posterior.

  Nicholls & Carmona (2020). *Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components.*

  Nicholls, Lee, Wu & Carmona (2022). *Valid belief updates for prequentially additive loss functions arising in Semi-Modular Inference.*

Variants

- Instead of completely cutting feedback, *semi-modular inference* controls the feedback, with a parameter in $[0, 1]$ to interpolate between the cut and the full posterior.

  Nicholls & Carmona (2020). *Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components.*

  Nicholls, Lee, Wu & Carmona (2022). *Valid belief updates for prequentially additive loss functions arising in Semi-Modular Inference.*

- Frazier & Nott (2022). *Cutting feedback and modularized analyses in generalized Bayesian inference.*

Variants

- Instead of completely cutting feedback, *semi-modular inference* controls the feedback, with a parameter in $[0, 1]$ to interpolate between the cut and the full posterior.

  Nicholls & Carmona (2020). *Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components.*

  Nicholls, Lee, Wu & Carmona (2022). *Valid belief updates for prequentially additive loss functions arising in Semi-Modular Inference.*

- Frazier & Nott (2022). *Cutting feedback and modularized analyses in generalized Bayesian inference.*

- Chakraborty, Nott, Drovandi, Frazier & Sisson (2022). *Modularized Bayesian analyses and cutting feedback in likelihood-free inference.*

Asymptotics

Asymptotics of two-step estimators in Murphy & Topel (1985).
Estimation and Inference in Two-Step Econometric Models.
Extended to cut distributions in Pompe & Jacob (2022).
Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap.

Asymptotics

Asymptotics of two-step estimators in Murphy & Topel (1985).
Estimation and Inference in Two-Step Econometric Models.
Extended to cut distributions in Pompe & Jacob (2022).
Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap.
Data generating process:

$$\text{scenario A} \quad p^\star(Y_{1,1:n_1}, Y_{2,1:n_2}) = \prod_{i=1}^{n_1} p_1^\star(Y_{1,i}) \prod_{i=1}^{n_2} p_2^\star(Y_{2,i}),$$

$$\text{scenario B} \quad p^\star(Y_{1,1:n_1}, Y_{2,1:n_2}) = \prod_{i=1}^{n} p^\star(Y_{1,i}, Y_{2,i}) \neq \prod_{i=1}^{n} p^\star(Y_{1,i}) p^\star(Y_{2,i}).$$

In scenario A assume $n_1/n_2 \to \alpha > 0$, and write $n = n_2$.
In scenario B $n_1 = n_2 = n$.

Asymptotics

Denote the two-step MLEs by $\hat{\theta}_1$ and $\hat{\theta}_2$, the latter given $\hat{\theta}_1$.

The asymptotic distribution of these MLEs is Normal with
variance Σ_A under scenario A,
variance Σ_B under scenario B.

Asymptotics

Denote the two-step MLEs by $\hat{\theta}_1$ and $\hat{\theta}_2$, the latter given $\hat{\theta}_1$.

The asymptotic distribution of these MLEs is Normal with
variance Σ_A under scenario A,
variance Σ_B under scenario B.

With (θ_1, θ_2) drawn from the cut distribution, under regularity conditions,

$$\sqrt{n}(\theta_1 - \hat{\theta}_1, \theta_2 - \hat{\theta}_2) \to \text{Normal}(0, \Sigma_C).$$

Interestingly Σ_C does not match Σ_A or Σ_B,
unless both models are well-specified AND scenario A holds.

Asymptotics

Frazier & Nott (2022). *Cutting feedback and modularized analyses in generalized Bayesian inference.*

Focus on the asymptotic behaviour of $p_2(\theta_2|\theta_1, Y_2)$,
assuming θ_1 is fixed in a neighborhood of its limit ${\theta_1}^\star$.

Asymptotics

Frazier & Nott (2022). *Cutting feedback and modularized analyses in generalized Bayesian inference.*

Focus on the asymptotic behaviour of $p_2(\theta_2|\theta_1, Y_2)$,
assuming θ_1 is fixed in a neighborhood of its limit ${\theta_1}^\star$.

The conditional distribution becomes Normal
with both mean and variance depending explicitly on θ_1.

Allows a finer understanding of the impact of θ_1 onto θ_2.

Supporters say aye, opponents say no

Setup: model 2 depends on an input that is itself estimated using model 1.

Cut distribution: obtain distribution of input using model 1 only, then estimate model 2 using a distribution of input.

- Can mitigate effect of misspecification.
- Facilitates interoperability across teams.
- Can resolve computational intractability of joint model.

Supporters say aye, opponents say no

Setup: model 2 depends on an input that is itself estimated using model 1.

Cut distribution: obtain distribution of input using model 1 only, then estimate model 2 using a distribution of input.

- ☺ Can mitigate effect of misspecification.
- ☺ Facilitates interoperability across teams.
- ☺ Can resolve computational intractability of joint model.

- ☹ Can lead to sub-optimal estimation/prediction since all data are not fully used.
- ☹ *In my experience, instead of "cutting", it works better to expand the model until it fits both datasets.*[1]
- ☹ Computation presents its own challenges.

[1] Gelman (2016) blog post entitled *Don't get me started on 'cut'.*

Asking the data whether to cut

Jacob, Murray, Holmes & Robert (2017). *Better together? Statistical learning in models made of modules.*

We can try to be principled about whether to cut or not.

Natural route: introduce measures of predictive performance that can be evaluated on test data.

Postulate: a distribution on parameters is good if it leads to accurate predictions.

Asking the data whether to cut

In the first module, θ_1 is defined in its relation to Y_1.

We propose to assess candidate distributions for θ_1 based on predictive performance for Y_1.

To decide between $p_1(\theta_1|Y_1)$ and $\pi(\theta_1|Y_1, Y_2)$, using the prequential approach and the logarithmic scoring rule, we compare $p_1(Y_1)$ with $\pi(Y_1|Y_2)$.

If $p_1(Y_1) > \pi(Y_1|Y_2)$, we support the use of distributions on (θ_1, θ_2) that admit $p_1(\theta_1|Y_1)$ as first marginal, e.g. cut.

Outline

Confusion about cut distributions

From Gelman (2020) blog post entitled *How to "cut" using Stan, if you must.*

Question (rephrased for brevity):

Have cut posteriors been implemented in Stan?

Reply:

> *This topic has come up before, and I don't think this "cut" is a good idea. If you want to implement it, [...] you'd first fit model 1 and get posterior simulations, then approx those simulations by a mixture of multivariate normal or t distributions, then use that as a prior for model 2. [...]*

Confusion about cut distributions

From Gelman (2020) blog post entitled *How to "cut" using Stan, if you must.*

Question (rephrased for brevity):

Have cut posteriors been implemented in Stan?

Reply:

> *This topic has come up before, and I don't think this "cut" is a good idea. If you want to implement it, [...] you'd first fit model 1 and get posterior simulations, then approx those simulations by a mixture of multivariate normal or t distributions, then use that as a prior for model 2. [...]*

This would in fact amount to a two-step approximation of the *standard* posterior distribution.

Modular approximations of the standard posterior

Lunn, Barrett, Sweeting & Thompson (2013). *Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis.*

Goudie, Presanis, Lunn, De Angelis & Wernisch (2016). *Model surgery: joining and splitting models with Markov melding.*

Manderson & Goudie (2021). *A numerically stable algorithm for integrating Bayesian models using Markov melding.*

Leonelli, Barons & Smith (2018). *A conditional independence framework for coherent modularized inference.*

Huge interest in approximating the *supraBayesian* with a de-centralized strategy, but this is not about cutting feedback.

Sampling from the cut distribution

The density of the cut distribution is

$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) \propto p_1(\theta_1) p_1(Y_1|\theta_1) \frac{p_2(\theta_2|\theta_1) p_2(Y_2|\theta_1, \theta_2)}{p_2(Y_2|\theta_1)}$$

$$\propto \frac{\pi(\theta_1, \theta_2|Y_1, Y_2)}{p_2(Y_2|\theta_1)}.$$

Sampling from the cut distribution

The density of the cut distribution is

$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) \propto p_1(\theta_1) p_1(Y_1|\theta_1) \frac{p_2(\theta_2|\theta_1) p_2(Y_2|\theta_1, \theta_2)}{p_2(Y_2|\theta_1)}$$
$$\propto \frac{\pi(\theta_1, \theta_2|Y_1, Y_2)}{p_2(Y_2|\theta_1)}.$$

The term $p_2(Y_2|\theta_1)$ is typically intractable,

$$p_2(Y_2|\theta_1) = \int p_2(Y_2|\theta_1, \theta_2) p_2(\theta_2|\theta_1) d\theta_2.$$

MCMC approach for *doubly intractable* targets:
Liu & Goudie (2021). *Stochastic approximation cut algorithm for inference in modularized Bayesian models.*

Sampling from the cut distribution

WinBUGS' approach via the `cut` function: alternate between

- sampling $\theta_1{}'$ from $K^1(\theta_1, d\theta_1{}')$ targeting $p_1(d\theta_1|Y_1)$,
- sampling $\theta_2{}'$ from $K^2_{\theta_1{}'}(\theta_2, d\theta_2{}')$ targeting $p_2(d\theta_2|\theta_1{}', Y_2)$.

This does not leave the cut distribution invariant. Iterating the kernel $K^2_{\theta_1'}$ enough times mitigates the issue.

Plummer (2014). *Cuts in Bayesian graphical models.*

Sampling from the cut distribution

In a *perfect sampling* world, we could sample

- θ_1 from $p_1(\theta_1|Y_1)$,
- θ_2 given θ_1 from $p_2(\theta_2|\theta_1, Y_2)$,

then (θ_1, θ_2) would be exactly following the cut distribution.

For many models, exact sampling is not feasible.

Sampling from the cut distribution

In an MCMC world, we can sample

- θ_1 approximately from $p_1(\theta_1|Y_1)$ using MCMC,
- θ_2 given θ_1 approximately from $p_2(\theta_2|\theta_1, Y_2)$ using MCMC,

then resulting samples approximate the cut distribution,
in the limit of the numbers of iterations in both stages.

Unbiased estimation of the cut distribution

By coupling pairs of MCMC chains, we can produce (random) empirical measures

$$\hat{\pi}(d\theta) = \sum_{n=1}^{N} \omega_n \delta_{\theta_n}(d\theta)$$

that approximate a target π, in the sense

$$\mathbb{E}\left[\sum_{n=1}^{N} \omega_n h(\theta_n)\right] = \int h(\theta)\pi(d\theta),$$

for a class of test functions h.

Jacob, O'Leary & Atchadé (2020). *Unbiased MCMC with couplings.*
Douc, Jacob, Lee & Vats (2022). *Solving the Poisson equation using coupled Markov chains.*

Unbiased estimation of the cut distribution

In an *unbiased MCMC* world, we can approximate

- $p_1(d\theta_1|Y_1)$ with a random measure

$$\hat{\pi}_1(d\theta_1) = \sum_{n=1}^{N_1} \omega_n \delta_{\theta_{1,n}}(d\theta_1),$$

- $p_2(d\theta_2|\theta_1, Y_2)$ for any θ_1, with a random measure

$$\hat{\pi}_2(d\theta_2|\theta_1) = \sum_{n=1}^{N_2} u_n \delta_{\theta_{2,n}}(d\theta_2).$$

Using the tower property, we can unbiasedly estimate

$$\int h(\theta_1, \theta_2) p_1(d\theta_1|Y_1) p_2(d\theta_2|\theta_1, Y_1).$$

Discussion

A very appealing aspect of Bayesian analysis is its unified treatment of many statistical questions.

Modular approaches appear to depart from the main framework, and thus bring discomfort.

Discussion

A very appealing aspect of Bayesian analysis is its unified treatment of many statistical questions.

Modular approaches appear to depart from the main framework, and thus bring discomfort.

If we consider the essence of Bayesian analysis:

- probability distributions for the quantities of interest,
- solid decision-theoretic foundations,
- wide applicability thanks to diverse computational strategies,

then modular approaches might still be essentially Bayesian.

Thank you!