Bayesian inference with models made of modules

Pierre E. Jacob

ESSEC Business School

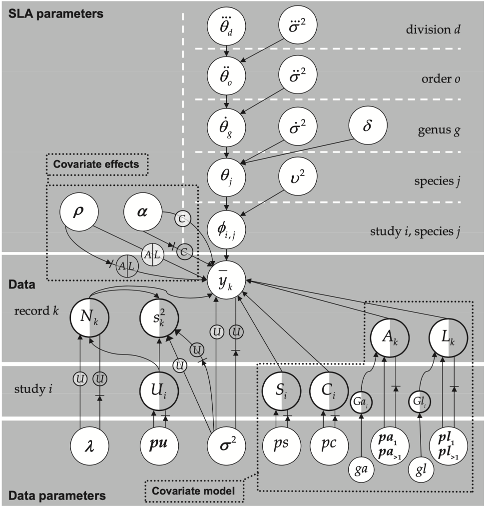
2022 ISBA World Meeting
Susie Bayarri Lecture
June 30, 2022

Outline

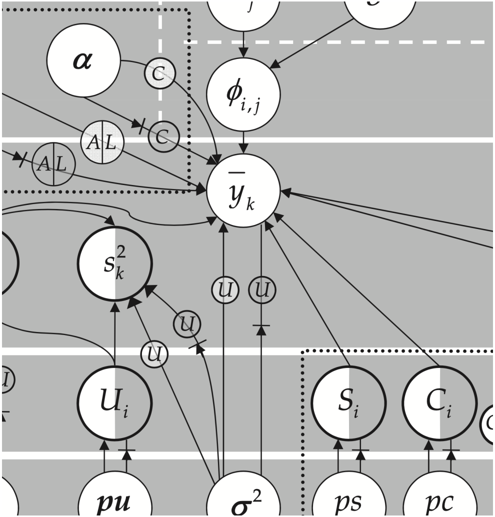
Outline

Let's start with a simple example…

Models made of modules



Ogle, Barber & Sartor (2013). *Feedback and Modularization in a Bayesian Meta–analysis of Tree Traits Affecting Forest Dynamics.*



Zooming in. . . we see arrows. . . and diodes ▶| . . . ?

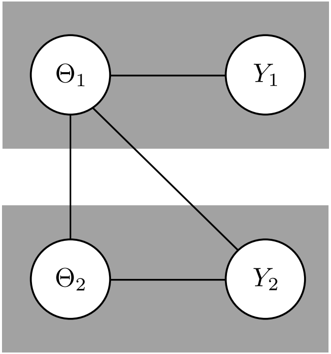
Models made of modules

- 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: $p_2\left(Y_2|\theta_1, \theta_2\right)$



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

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

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

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

Departures from the joint model approach: why, how, when?

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: SARS-COV-2 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 $\mathbb{P}(\text{tested}|\text{not infected})$.

Example: SARS-COV-2 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 $\mathbb{P}(\text{tested}|\text{not infected})$.

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: stochastic dynamical 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: stochastic dynamical 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?

Example: PKPD

Bennett & Wakefield (2001). *Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling.*

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: HPV prevalence and cervical cancer incidence

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:

$$\forall i = 1, \ldots, I \quad Y_i \sim \text{Poisson}(\lambda_i T_i), \quad \log(\lambda_i) = \eta_1 + \eta_2\, \varphi_i,$$

Y_i is number of cases during study in country i,
T_i: woman-years of follow-up in country i.

Example: two-step estimation

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

Impact of unanticipated money growth on unemployment.

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

  M_t: proportional growth in the M1 definition of money,

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

- $$\forall t \quad \log \frac{\mathrm{U}_t}{1 - \mathrm{U}_t} = \beta X_{2t} + \gamma\, \epsilon_t + W_t,$$

  U_t: annual average unemployment rate,

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

Example: two-step estimation

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

Impact of unanticipated money growth on unemployment.

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

  M_t: proportional growth in the M1 definition of money,

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

- $$\forall t \quad \log \frac{\mathrm{U}_t}{1 - \mathrm{U}_t} = \beta X_{2t} + \gamma\, \epsilon_t + W_t,$$

  U_t: annual average unemployment rate,

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

Joint estimation "inappropriate or computationally infeasible".

Example: multiple imputation

- Missing data: imputation of missing values, then analysis of completed data.

  Jackson, Best & Richardson (2009). *Bayesian graphical models for regression on multiple data sets with different variables.*

Example: multiple imputation

- Missing data: imputation of missing values, then analysis of completed data.

  Jackson, Best & Richardson (2009). *Bayesian graphical models for regression on multiple data sets with different variables.*

- Multiphase inference: a first analyst pre-processes raw data, then a second analyst uses the processed data.

  Blocker & Meng (2013). *The potential and perils of preprocessing: Building new foundations.*

More examples

- Environmental epidemiology: estimation of environmental exposure, then associated health effects.

  Blangiardo, Hansell & Richardson (2011). *A Bayesian model of time activity data to investigate health effect of air pollution in time series studies.*

More examples

- Environmental epidemiology: estimation of environmental exposure, then associated health effects.

  Blangiardo, Hansell & Richardson (2011). *A Bayesian model of time activity data to investigate health effect of air pollution in time series studies.*

- Causal inference with propensity scores: estimation of probability of individuals receiving treatment, then treatment effect adjusted for propensity score.

  Zigler, Watts, Yeh, Wang, Coull & Dominici (2013). *Model feedback in Bayesian propensity score estimation.*

  Saarela, Belzile & Stephens (2016). *A Bayesian view of doubly robust causal inference.*

More examples

- Environmental epidemiology: estimation of environmental exposure, then associated health effects.

  Blangiardo, Hansell & Richardson (2011). *A Bayesian model of time activity data to investigate health effect of air pollution in time series studies.*

- Causal inference with propensity scores: estimation of probability of individuals receiving treatment, then treatment effect adjusted for propensity score.

  Zigler, Watts, Yeh, Wang, Coull & Dominici (2013). *Model feedback in Bayesian propensity score estimation.*

  Saarela, Belzile & Stephens (2016). *A Bayesian view of doubly robust causal inference.*

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. first estimate θ_1 given Y_1, 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. first estimate θ_1 given Y_1, 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.
- ☺ Misspecification of 2nd module does not impact θ_1.

Cut approach

Propagate uncertainty without allowing feedback.

Define the cut distribution:

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

Cut approach

Propagate uncertainty without allowing feedback.

Define the cut distribution:

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

Known as Bayesian two-step estimation, or as "cutting feedback", term suggested by Nicky Best according to Jonty Rougier, in a comment on Sansó, Forest & Zantedeschi (2008).

Cut approach

Propagate uncertainty without allowing feedback.

Define the cut distribution:

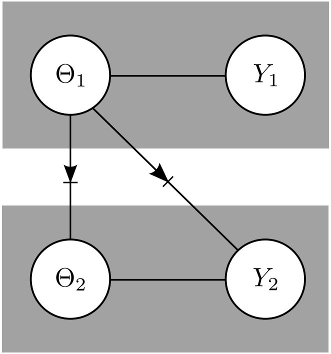
$$\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2) = p_1(\theta_1 | Y_1) p_2(\theta_2 | \theta_1, Y_2).$$

Known as Bayesian two-step estimation, or as "cutting feedback", term suggested by Nicky Best according to Jonty Rougier, in a comment on Sansó, Forest & Zantedeschi (2008).

Ideal sampling procedure:

1. Sample θ_1 from $p_1(\theta_1 | Y_1)$.
2. Sample θ_2 given θ_1 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 density functions:

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

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

Cut approach

Difference between cut and standard posterior density functions:

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

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

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 density functions:

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

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

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

The marginal distribution of θ_1 differs,

$$p_1(\theta_1|Y_1) \quad \text{for cut,} \quad \pi(\theta_1|Y_1, Y_2) \quad \text{for standard posterior,}$$

The conditional distribution of θ_2 is the same: $p_2(\theta_2|\theta_1, Y_2)$.

A variational representation

Among all distributions $q(\theta_1, \theta_2)$ with marginal $p_1(\theta_1|Y_1)$ on θ_1, $\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2)$ minimizes

$$\text{Kullback–Leibler}\left(q(\theta_1, \theta_2), \pi(\theta_1, \theta_2 | Y_1, Y_2)\right).$$

Lemma 1 in Yu, Nott & Smith (2021). *Variational inference for cutting feedback in misspecified models.*

A variational representation

Among all distributions $q(\theta_1, \theta_2)$ with marginal $p_1(\theta_1|Y_1)$ on θ_1, $\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2)$ minimizes

$$\text{Kullback–Leibler}\left(q(\theta_1, \theta_2), \pi(\theta_1, \theta_2|Y_1, Y_2)\right).$$

Lemma 1 in Yu, Nott & Smith (2021). *Variational inference for cutting feedback in misspecified models.*

We can also view the cut distribution as a valid representation of beliefs about the parameters.

Updating beliefs

Bissiri, Holmes & Walker (2016). *A general framework for updating belief distribution.*

Prior beliefs $p(\theta)$, loss associated with data $l(y, \theta)$,
assume loss and prior are independent pieces of information.

Updating beliefs

Bissiri, Holmes & Walker (2016). *A general framework for updating belief distribution.*

Prior beliefs $p(\theta)$, loss associated with data $l(y,\theta)$,
assume loss and prior are independent pieces of information.

We look for an update "$p(\theta|y)$"$= \psi(l(y,\theta), p(\theta))$.

Updating beliefs

Bissiri, Holmes & Walker (2016). *A general framework for updating belief distribution.*

Prior beliefs $p(\theta)$, loss associated with data $l(y, \theta)$,
assume loss and prior are independent pieces of information.

We look for an update "$p(\theta|y)$"$= \psi(l(y, \theta), p(\theta))$.

Coherence:

$$\psi(l(y', \theta), \psi(l(y, \theta), p(\theta))) = \psi(l(y', \theta) + l(y, \theta), p(\theta)).$$

Updating beliefs

Bissiri, Holmes & Walker (2016). *A general framework for updating belief distribution.*

Prior beliefs $p(\theta)$, loss associated with data $l(y, \theta)$,
assume loss and prior are independent pieces of information.

We look for an update "$p(\theta|y)$"$= \psi(l(y, \theta), p(\theta))$.

Coherence:

$$\psi(l(y', \theta), \psi(l(y, \theta), p(\theta))) = \psi(l(y', \theta) + l(y, \theta), p(\theta)).$$

Result: for coherence and optimality, we need to define

$$p(\theta|y) = \text{argmin}_q \int l(y, \theta) q(d\theta) + \text{KL}(q(\theta), p(\theta)).$$

Solution: $p(\theta|y) \propto \exp(-l(y, \theta))p(\theta)$.

Cut distributions are valid belief updates

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

Retrieve the cut distribution in the framework of Bissiri, Holmes & Walker (2016), with the loss

$$\text{“}l(y,\theta)\text{”} = -\left\{\log p(Y_1|\theta_1)p(Y_2|\theta_1,\theta_2) - \log p_2(Y_2|\theta_1)\right\}.$$

Cut distributions are valid belief updates

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

Retrieve the cut distribution in the framework of Bissiri, Holmes & Walker (2016), with the loss

$$\text{“}l(y,\theta)\text{”} = -\left\{\log p(Y_1|\theta_1)p(Y_2|\theta_1,\theta_2) - \log p_2(Y_2|\theta_1)\right\}.$$

The loss involves $p_2(Y_2|\theta_1) = \int p_2(Y_2|\theta_1,\theta_2)p_2(\theta_2|\theta_1)d\theta_2$ and thus depends on the prior... the argument needs adjustments.

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

Variants

- Allow a controlled amount of feedback.

  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

- Allow a controlled amount of feedback.

  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.*

- Cut + replace minus log-likelihood by other loss functions.

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

scenario A $\quad n_1/n_2 \to \alpha > 0$

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

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.

scenario A $\quad n_1/n_2 \to \alpha > 0$

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

scenario B $\quad n_1 = n_2 = n$

$$\begin{aligned} p^\star(Y_{1,1:n_1}, Y_{2,1:n_2}) &= \prod_{i=1}^{n} p^\star(Y_{1,i}, Y_{2,i}) \\ &\neq \prod_{i=1}^{n} p_1^\star(Y_{1,i}) p_2^\star(Y_{2,i}) \end{aligned}$$

Asymptotics

Under regularity conditions, introduce the two-step MLEs:

$$\hat{\theta}_1 = \text{argmax}_{\theta_1} \log p_1(Y_1|\theta_1) \longrightarrow \theta_1^\star,$$
$$\hat{\theta}_2 = \text{argmax}_{\theta_2} \log p_2(Y_2|\hat{\theta}_1, \theta_2) \longrightarrow \theta_2^\star.$$

Their asymptotic joint distribution is Normal with variance Σ_A under scenario A, Σ_B under scenario B.

Asymptotics

Under regularity conditions, introduce the two-step MLEs:

$$\hat{\theta}_1 = \text{argmax}_{\theta_1} \log p_1(Y_1|\theta_1) \longrightarrow \theta_1^\star,$$
$$\hat{\theta}_2 = \text{argmax}_{\theta_2} \log p_2(Y_2|\hat{\theta}_1, \theta_2) \longrightarrow \theta_2^\star.$$

Their asymptotic joint distribution is Normal with variance Σ_A under scenario A, Σ_B under scenario B.

Sandwich formula: $\mathbb{E}^\star[-\frac{d^2\ell(\theta^\star)}{d\theta^2}]^{-1}\mathbb{E}^\star[(\frac{d\ell(\theta^\star)}{d\theta})^2]\mathbb{E}^\star[-\frac{d^2\ell(\theta^\star)}{d\theta^2}]^{-1}$
+ possible dependencies between Y_1 and Y_2 in scenario B.

Asymptotics

Under regularity conditions, introduce the two-step MLEs:

$$\hat{\theta}_1 = \text{argmax}_{\theta_1} \log p_1(Y_1|\theta_1) \longrightarrow \theta_1^\star,$$
$$\hat{\theta}_2 = \text{argmax}_{\theta_2} \log p_2(Y_2|\hat{\theta}_1, \theta_2) \longrightarrow \theta_2^\star.$$

Their asymptotic joint distribution is Normal with variance Σ_A under scenario A, Σ_B under scenario B.

Sandwich formula: $\mathbb{E}^\star[-\frac{d^2\ell(\theta^\star)}{d\theta^2}]^{-1}\mathbb{E}^\star[(\frac{d\ell(\theta^\star)}{d\theta})^2]\mathbb{E}^\star[-\frac{d^2\ell(\theta^\star)}{d\theta^2}]^{-1}$
$+$ possible dependencies between Y_1 and Y_2 in scenario B.

For (θ_1, θ_2) drawn from the cut distribution,

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

$\Sigma_C \equiv \mathbb{E}^\star[-\frac{d^2\ell(\theta^\star)}{d\theta^2}]^{-1}$ does not match Σ_A or Σ_B.

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 the MLE 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 the MLE limit $\theta_1^\star$.

The asymptotic conditional distribution $p_2(\theta_2|\theta_1, Y_2)$ is Normal with both mean and variance depending explicitly on θ_1.

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

Supporters say aye, opponents say no

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

Cut distribution:

- ☺ Can mitigate effect of misspecification.
- ☺ Facilitates interoperability across teams.
- ☺ Can resolve computational intractability of joint model.
- ☺ Not completely unprincipled.

Supporters say aye, opponents say no

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

Cut distribution:

- ☺ Can mitigate effect of misspecification.
- ☺ Facilitates interoperability across teams.
- ☺ Can resolve computational intractability of joint model.
- ☺ Not completely unprincipled.

- ☹ Can lead to sub-optimal estimation/prediction accuracy.
- ☹ Is no replacement for constructive model criticism.
- ☹ Associated computations present their own challenges.

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. But which data: Y_1? Y_2?

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. But which data: Y_1? Y_2?

Postulate: parameters are meaningful in the module that first defines them. Thus distributions of parameters should be compared on predictions in the module that defines them.

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.

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.

Two candidates:

$$p_1(\theta_1|Y_1) \quad \text{and} \quad \pi(\theta_1|Y_1, Y_2).$$

Using the prequential approach and the logarithmic scoring rule, we compare

$$p_1(Y_1) \quad \text{and} \quad \pi(Y_1|Y_2).$$

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.

Two candidates:

$$p_1(\theta_1|Y_1) \quad \text{and} \quad \pi(\theta_1|Y_1, Y_2).$$

Using the prequential approach and the logarithmic scoring rule, we compare

$$p_1(Y_1) \quad \text{and} \quad \pi(Y_1|Y_2).$$

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

1 Models made of modules and issues with joint modeling

2 Cutting feedback: cut posterior distributions

3 Computation involved when cutting feedback

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?

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, not the cut 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.*

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

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

Sampling from the cut distribution

The density of the cut distribution is

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

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

OpenBUGS' 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,

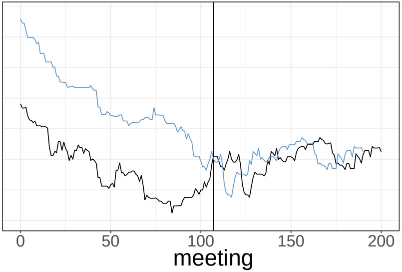
then resulting samples approximate the cut distribution,
in the limit of the numbers of iterations in both stages.

☹ Can involve tuning and convergence diagnostics for many MCMC runs at the 2nd stage, each with a different target.

Unbiased estimation of the cut distribution

Jacob, O'Leary & Atchadé (2020). *Unbiased Markov chain Monte Carlo with couplings.*

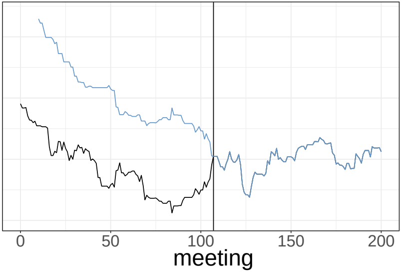
By coupling pairs of π-invariant chains in a particular way,



Unbiased estimation of the cut distribution

Jacob, O'Leary & Atchadé (2020). *Unbiased Markov chain Monte Carlo with couplings.*

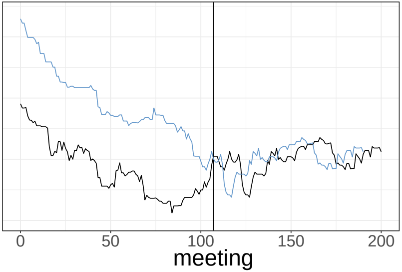
By coupling pairs of π-invariant chains in a particular way,



Unbiased estimation of the cut distribution

Jacob, O'Leary & Atchadé (2020). *Unbiased Markov chain Monte Carlo with couplings.*

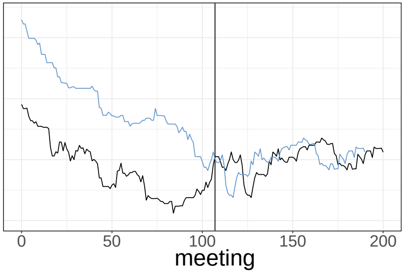
By coupling pairs of π-invariant chains in a particular way,



Unbiased estimation of the cut distribution

Jacob, O'Leary & Atchadé (2020). *Unbiased Markov chain Monte Carlo with couplings.*

By coupling pairs of π-invariant chains in a particular way,



we can construct a signed measure $\hat{\pi}(d\theta) = \sum_{n=1}^{N} \omega_n \delta_{\theta_n}(d\theta)$ with $\mathbb{E}[\hat{\pi}(h)] = \pi(h)$ for a class of test functions h.

Unbiased estimation of the cut distribution

In an *unbiased MCMC* world, we can approximate without bias

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

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

obtained from coupled $p_1(d\theta_1|Y_1)$-invariant chains.

Unbiased estimation of the cut distribution

In an *unbiased MCMC* world, we can approximate without bias

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

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

  obtained from coupled $p_1(d\theta_1|Y_1)$-invariant chains.

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

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

  obtained from coupled $p_2(d\theta_2|\theta_1, Y_2)$-invariant chains.

Unbiased estimation of the cut distribution

In an *unbiased MCMC* world, we can approximate without bias

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

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

  obtained from coupled $p_1(d\theta_1|Y_1)$-invariant chains.

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

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

  obtained from coupled $p_2(d\theta_2|\theta_1, Y_2)$-invariant chains.

Using the tower property, we can estimate without bias expectations with respect to $\pi^{\text{cut}}(\theta_1, \theta_2; Y_1, Y_2)$.

Inexact approximations of the cut distribution

- Variational inference:

  Yu, Nott & Smith (2021). *Variational inference for cutting feedback in misspecified models.*

  Carmona & Nicholls (2022). *Scalable Semi-Modular Inference with Variational Meta-Posteriors.*

Inexact approximations of the cut distribution

- Variational inference:

  Yu, Nott & Smith (2021). *Variational inference for cutting feedback in misspecified models.*

  Carmona & Nicholls (2022). *Scalable Semi-Modular Inference with Variational Meta-Posteriors.*

- Posterior bootstrap:

  Pompe & Jacob (2022). *Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap.*

  adapting techniques developed earlier

  by Newton & Raftery, Fong, Lyddon, Holmes & Walker.

Inexact approximations of the cut distribution

- Variational inference:

  Yu, Nott & Smith (2021). *Variational inference for cutting feedback in misspecified models.*

  Carmona & Nicholls (2022). *Scalable Semi-Modular Inference with Variational Meta-Posteriors.*

- Posterior bootstrap:

  Pompe & Jacob (2022). *Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap.*

  adapting techniques developed earlier

  by Newton & Raftery, Fong, Lyddon, Holmes & Walker.

- Modular approximate Bayesian computation

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

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 the *essence* of Bayesian analysis is not the direct application of Bayes' theorem, but

- probability distributions for the quantities of interest, constructed from data and prior information,
- a framework for statistical inference supported by principles & decision theory,

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 the *essence* of Bayesian analysis is not the direct application of Bayes' theorem, but

- probability distributions for the quantities of interest, constructed from data and prior information,
- a framework for statistical inference supported by principles & decision theory,
- a good excuse to play with Markov chains,

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 the *essence* of Bayesian analysis is not the direct application of Bayes' theorem, but

- probability distributions for the quantities of interest, constructed from data and prior information,
- a framework for statistical inference supported by principles & decision theory,
- a good excuse to play with Markov chains,

then cut posteriors might be *essentially* Bayesian.

Thank you!