Semi-Modular Inference
Supplementary material

Chris U. Carmona
Department of Statistics
University of Oxford
Oxford, U
carmona@stats.ox.ac.uk

Geoff K. Nicholls
Department of Statistics
University of Oxford
Oxford, UK
nicholls@stats.ox.ac.uk

In this document we present additional material which complements the main article. The section numbers in this document are aligned with the sections in the main article, for the ease of referencing.

Notation: For the sake of compactness in our derivations, the following expressions are used:

$$\begin{aligned}
p(Z) &= \int\int p(Z \mid \varphi)\, p(\varphi)\, d\varphi \\
p(Z,Y)_\theta &= \int\int p(Z \mid \varphi)\, p(Y \mid \varphi,\theta)\, p(\varphi,\theta)\, d\varphi\, d\theta \\
p(Z,Y_\eta)_{\tilde\theta} &= \int\int p(Z \mid \varphi)\, p(Y \mid \varphi,\tilde\theta)^\eta\, p(\varphi,\tilde\theta)\, d\varphi\, d\tilde\theta \\
p(Y_\eta)_{\tilde\theta} &= \int\int p(Y \mid \varphi,\tilde\theta)^\eta\, p(\varphi,\tilde\theta)\, d\varphi\, d\tilde\theta \\
p(Y,\varphi)_\theta &= \int p(Y \mid \varphi,\theta)\, p(\varphi,\theta)\, d\theta \\
p(Y \mid \varphi)_\theta &= \frac{1}{p(\varphi)}\, p(Y,\varphi)_\theta
\end{aligned}$$

All the figures and numerical results presented in the main text and this supplementcan be replicated using our R package, `aistats2020smi`, available in Github.

```
> # devtools::install_github("christianu7/aistats2020smi")
> library(aistats2020smi)
```

1 Introduction

Here are a few remarks on model misspecification. Classically, misspecification is identified in goodness-of-fit checks as poor posterior predictive performance on held out data. In this paper, a model is relatively more misspecified if it has relatively worse performance in posterior predictive checks. Notice that this may be caused by a misspecified observation model, as in the HPV example in the main text, but unrepresentative prior assumptions may also lead to mispecification. This is illustrated in Section 5.1. The example in Section 5.2 arguably suffers from both forms of mispecification.

The SMI procedure can alternatively be thought of as measuring misspecification. We can measure a model's goodness of fit using the η^*-values of its modules. Using the degree of influence as a measure of misfit has the advantage that it is defined on a standard scale $[0,1]$ with $\eta^*=1$ corresponding to no evidence for misfit and $\eta^*=0$ indicating the module should be removed entirely when we estimate parameters in the other modules, and indicating substantial misspecification.

2 Background methods

2.1 Modular Inference: cut model

Explicit formulae for cut posterior

The graphical model analysed in the main text is shown in Figure 1.

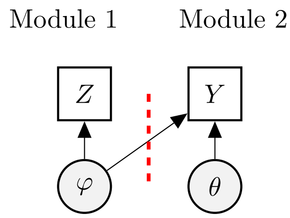

Figure 1: Graphical representation of a simple multi-modular model.

The **conventional (full) posterior** for this model is:

$$\begin{aligned}
p(\varphi, \theta \mid Z, Y) &= p(\varphi \mid Z, Y)\; p(\theta \mid Y, \varphi) \\
&= p(Z, Y, \varphi, \theta)\; \frac{1}{p(Z, Y)_\theta} \\
&= p(Z \mid \varphi)\; p(Y \mid \varphi, \theta)\; p(\varphi, \theta)\; \frac{1}{p(Z, Y)_\theta} \qquad (1)
\end{aligned}$$

The **cut posterior** for this model is defined (Plummer, 2015) as:

$$\begin{aligned}
p_{cut}(\varphi, \theta \mid Z, Y) &= p(\varphi \mid Z)\; p(\theta \mid Y, \varphi) \\
&= \frac{p(Z \mid \varphi) p(\varphi)}{p(Z)} \frac{p(Y \mid \varphi, \theta)\; p(\varphi, \theta)}{p(Y, \varphi)_\theta} \\
&= p(Z \mid \varphi)\; p(Y \mid \varphi, \theta)\; \frac{1}{p(Z)}\; \frac{1}{p(Y \mid \varphi)_\theta} p(\varphi, \theta) \qquad (2)
\end{aligned}$$

Note the relation between the cut posterior and the conventional posterior

$$\begin{aligned}
p_{cut}(\varphi, \theta \mid Z, Y) &= p(Z, Y, \varphi, \theta)\; \frac{1}{p(Z)\; p(Y \mid \varphi)_\theta} \\
&= p(\varphi, \theta \mid Z, Y)\; \frac{p(Z, Y)}{p(Z)\; p(Y \mid \varphi)_\theta}
\end{aligned}$$

3 Semi-Modular Inference

3.1 SMI distributions

Explicit formulae for SMI posterior

The η**-smi posterior** is defined as

$$\begin{aligned}
p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) &= p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)\; p(\theta \mid Y, \varphi) \\
&= \frac{p(Z \mid \varphi)\; p(Y \mid \varphi, \tilde{\theta})^\eta\; p(\varphi, \tilde{\theta})}{p(Z, Y)_{\tilde{\theta}}} \frac{p(Y \mid \varphi, \theta)\; p(\varphi, \theta)}{p(Y, \varphi)_\theta} \\
&= p(Z \mid \varphi)\; p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta)\; \frac{1}{p(Z, Y_\eta)_{\tilde{\theta}}\; p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}) \qquad (3) \\
&\propto p(Z \mid \varphi)\; p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta)\; \frac{1}{p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta})
\end{aligned}$$

In the penultimate step, we assume that θ and $\tilde{\theta}$ are conditionally independent given φ in the prior, so $p(\varphi, \theta, \tilde{\theta}) = p(\theta \mid \varphi)p(\tilde{\theta} \mid \varphi)p(\varphi) = p(\varphi, \theta)\ p(\varphi, \tilde{\theta})\frac{1}{p(\varphi)}$.

From here is easy to see two particular cases of the η-smi posterior: taking marginals over $\tilde{\theta}$ in SMI, the cut model $p_{smi,\eta}(\varphi, \theta \mid Z, Y) = p_{cut}(\varphi, \theta \mid Z, Y)$ when $\eta = 0$; and the conventional posterior $p_{smi,\eta}(\varphi, \theta \mid Z, Y) = p(\varphi, \theta \mid Z, Y)$ when $\eta = 1$.

4 Analysis with (Semi-)Modular Inference

4.1 Coherence of (Semi-)Modular Inference

4.1.1 Coherent update of beliefs

In Bissiri et al. (2016), the conventional update of beliefs provided by Bayes theorem is expanded, providing a generalised framework in which alternative inferential options are justified beyond the conventional posterior.

Let's use single-module notation in this Subsection 4.1.1, so we are aligned with Bissiri et al. (2016). Denote newly observe data as Y and the parameter of interest θ.

The framework established by Bissiri et al. (2016) relies on the idea that an update of beliefs must exist. This update of beliefs is performed, under a decision theory framework, by a function ψ, which turns the prior into posterior beliefs by incorporating new observed data y via a loss function $l(\theta; y)$. that is:

$$p(\theta \mid y) = \psi\{l(\theta; y), p(\theta)\}$$

Such update ψ is **Coherent** if it ensures that we preserve the posterior regardless of the order in which the data was observed. In other words, the posterior is the same whether we update our beliefs by observing all data simultaneously or by observing the data sequentially,

$$\psi\{l(\theta; x_2), \psi\{l(\theta; x_1), p(\theta)\}\} = \psi\{l(\theta; x_1) + l(\theta; x_2), p(\theta)\}$$

The authors show that an optimal, valid and coherent update of beliefs is of the form

$$p(\theta \mid y) = \psi\{l(\theta; y), p(\theta)\} = \frac{\exp\{-l(\theta; y)\}\ p(\theta)}{\int \exp\{-l(\theta; y)\}\ p(\theta)\ d\theta}$$

4.1.2 Coherence of the SMI posterior

Back in our multi-modular setting, the flexibility of the framework allows us to analyse the SMI posterior and the cut model posterior as valid schemes for the update of beliefs, which differ from the traditional fully-Bayesian update.

In the proofs which follow we work with SMI at an arbitrary fixed η. Since this means the results hold at $\eta = 0$, they hold for modular inference/cut models/Bayesian multiple imputation, at least for imputation in the modular setting we consider.

From Eq. 2 we can see that the loss function underlying the update of beliefs in the cut model is

$$l_{cut}((\varphi, \theta); (Z, Y)) = -\log p(Z \mid \varphi) - \log p(Y \mid \varphi, \theta) + \log p(Y \mid \varphi). \tag{4}$$

Similarly, from Eq. 3 we can derive the loss function underlying the update of beliefs in SMI the posterior,

$$l_{smi,\eta}((\varphi, \theta, \tilde{\theta}); (Z, Y)) = -\log p(Z \mid \varphi) - \eta \log p(Y \mid \varphi, \tilde{\theta}) - \log p(Y \mid \varphi, \theta) + \log p(Y \mid \varphi). \tag{5}$$

Here we prove that the SMI posterior preserves multi-modular coherence. In multi-modular settings, coherence must hold in two ways: 1) by observing responses from different modules one after the other (i.e. first Z, and then Y); and 2) by observing sequential fragments within the same module (e.g. first Z_1, and then Z_2, with $Z = (Z_1, Z_2)$).

The SMI posterior in Eq. 3 updates the belief distribution by observing the two datasets simultaneously. Under coherent inference, we can also update by observing data only from one module at a time, and still preserve the loss function in eq.5.

Say our current distribution of beliefs about $(\varphi, \theta, \tilde{\theta})$ is $p(\varphi, \theta, \tilde{\theta})$. Our updated belief by observing *only* Z would be

$$\begin{aligned} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z) &= \psi\{l((\varphi, \theta, \tilde{\theta}); Z), p(\varphi, \theta, \tilde{\theta})\} \\ &= p(Z \mid \varphi) \, \frac{1}{p(Z)} p(\varphi, \theta, \tilde{\theta}) \\ &= p(Z \mid \varphi) \frac{1}{\int p(Z \mid \varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}) \end{aligned} \tag{6}$$

similarly, if we observe *only* Y our updated beliefs are

$$\begin{aligned} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Y) &= \psi\{l((\varphi, \theta, \tilde{\theta}); Y), p(\varphi, \theta, \tilde{\theta})\} \\ &= p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta) \, \frac{1}{p(Y_\eta)_{\tilde{\theta}} \, p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}) \\ &= p(Y \mid \varphi, \tilde{\theta})^\eta \, p(Y \mid \varphi, \theta) \frac{1}{\int \int p(Y \mid \varphi, \tilde{\theta})^\eta \, p(\varphi, \tilde{\theta}) \, d\varphi \, d\tilde{\theta}} \frac{p(\varphi)}{\int p(Y \mid \varphi, \theta) \, p(\varphi, \theta) \, d\theta} p(\varphi, \theta, \tilde{\theta}) \end{aligned} \tag{7}$$

Coherence when observing data for different modules sequentially.

First, we show that the update from prior to posterior, is equivalent to updating sequentially first Z and then Y. This is (a)=(b)+(c) in the following diagram:

$$\begin{tikzcd} p(\varphi, \theta, \tilde{\theta}) \arrow[r, "(b)"'] \arrow[rr, bend left, "(a)"] & p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z) \arrow[r, "(c)"'] & p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) \end{tikzcd}$$

The update (a) is Eq. 3, updating beliefs by only observing both Z and Y

$$p_{(a)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) = p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)$$

The update (b) is similar to Eq. 6, updating beliefs only with data Z.

$$p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z) = p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z) \tag{8}$$

$$= p(Z \mid \varphi) \, \frac{1}{p(Z)} p(\varphi, \theta, \tilde{\theta})$$

$$= p(Z \mid \varphi) \, \frac{1}{\int p(Z \mid \varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}). \tag{9}$$

The update (c) is equivalent to Eq. 7 substituting the current beliefs with $p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z)$, and updating beliefs only with data Y.

$$\begin{aligned} p_{(b)+(c)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) &\propto p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta) \frac{p_{(b)}(\varphi \mid Z)}{\int p(Y \mid \varphi, \theta) \, p_{(b)}(\varphi, \theta \mid Z) \, d\theta} p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z) \\ &\propto p(Z \mid \varphi) p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta) \frac{1}{P(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}). \end{aligned}$$

The equivalence $p_{(b)+(c)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) = p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)$ is clear by comparing the last formula with smi posterior in Eq. 3. For the last line, we used the following identity

$$\begin{aligned} \frac{p_{(b)}(\varphi \mid Z)}{\int p(Y \mid \varphi, \theta) \, p_{(b)}(\varphi, \theta \mid Z) \, d\theta} &= \frac{\int \int p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z) d\theta d\tilde{\theta}}{\int \int p(Y \mid \varphi, \theta) \, p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z) \, d\theta d\tilde{\theta}} \\ &= \frac{\int \int p(Z \mid \varphi) \, \frac{1}{\int p(Z \mid \varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}) d\theta d\tilde{\theta}}{\int \int p(Y \mid \varphi, \theta) \, p(Z \mid \varphi) \, \frac{1}{\int p(Z \mid \varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}) \, d\theta d\tilde{\theta}} \\ &= \frac{p(\varphi)}{\int p(Y \mid \varphi, \theta) \, p(\varphi, \theta) \, d\theta} \\ &= \frac{1}{p(Y \mid \varphi)_\theta}. \end{aligned}$$

Now, we want a similar result by first observing Y and then Z, i.e. (a)=(d)+(e) in the following diagram

$$\begin{tikzcd} p(\varphi,\theta,\tilde{\theta}) \arrow[r, "(d)"'] \arrow[rr, bend left, "(a)"] & p_{smi,\eta}(\varphi,\theta,\tilde{\theta} \mid Y) \arrow[r, "(e)"'] & p_{smi,\eta}(\varphi,\theta,\tilde{\theta} \mid Z, Y) \end{tikzcd}$$

The update (a) is again given by Equation 3.

The update (d) is Eq. 7

$$p_{(d)}(\varphi,\theta,\tilde{\theta} \mid Y) \propto p(Y \mid \varphi,\tilde{\theta})^{\eta} p(Y \mid \varphi,\theta)\, \frac{1}{p(Y \mid \varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}). \tag{10}$$

The update (d)+(e) is equivalent to Eq. 6 substituting the current beliefs with $p_{(d)}(\varphi,\theta,\tilde{\theta} \mid Y)$

$$\begin{aligned} p_{(d)+(e)}(\varphi,\theta,\tilde{\theta} \mid Z, Y) &\propto p(Z \mid \varphi)\; p_{(d)}(\varphi,\theta,\tilde{\theta} \mid Y) \\ &= p(Z \mid \varphi)\; p(Y \mid \varphi,\tilde{\theta})^{\eta}\; p(Y \mid \varphi,\theta)\, \frac{1}{p(Y \mid \varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}). \end{aligned}$$

The equivalence $p_{(d)+(e)}(\varphi,\theta,\tilde{\theta} \mid Z, Y) = p_{smi,\eta}(\varphi,\theta,\tilde{\theta} \mid Z, Y)$ is direct from comparing the last formula with Eq. 3.

Coherence when observing data partitioned from the same module.

We now verify that the SMI posterior is coherent when observing a sequential portions of the same module. Define the partitions $Z = (Z_1, Z_2)$ and $Y = (Y_1, Y_2)$.

First, we verify coherence for the partition of data Z. We want to check (b)=(b1)+(b2) in the following diagram. $\begin{tikzcd} p(\varphi,\theta,\tilde{\theta}) \arrow[r, "(b1)"'] \arrow[rr, bend left, "(b)"] & p_{smi,\eta}(\varphi,\theta,\tilde{\theta} \mid Z_1) \arrow[r, "(b2)"'] & p_{smi,\eta}(\varphi,\theta,\tilde{\theta} \mid Z) \end{tikzcd}$

Update (b) is the same as defined above in Equation 8

Updates (b1) and (b2) are similar to Eq. 6, substituting the corresponding Z and current state of beliefs

$$\begin{aligned} p_{(b1)}(\varphi,\theta,\tilde{\theta} \mid Z_1) &= p(Z_1 \mid \varphi)\, \frac{1}{\int p(Z_1 \mid \varphi)\; p(\varphi)\; d\varphi} p(\varphi,\theta,\tilde{\theta}), \\ p_{(b1)+(b2)}(\varphi,\theta,\tilde{\theta} \mid Z_1, Z_2) &= p(Z_2 \mid \varphi)\, \frac{1}{\int p_{(b1)}(Z_2 \mid \varphi)\; p_{(b1)}(\varphi \mid Z_1)\; d\varphi} p_{(b1)}(\varphi,\theta,\tilde{\theta} \mid Z_1) \\ &\propto p(Z_1 \mid \varphi)\; p(Z_2 \mid \varphi)\; p(\varphi,\theta,\tilde{\theta}) \\ &= p(Z \mid \varphi)\; p(\varphi,\theta,\tilde{\theta}), \end{aligned}$$

clearly $p_{(b1)+(b2)}(\varphi,\theta,\tilde{\theta} \mid Z_1, Z_2) = p_{(b)}(\varphi,\theta,\tilde{\theta} \mid Z)$.

Lastly, we verify coherence for the partition of data Y. We want to check (d)=(d1)+(d2) in the following diagram.

$$\begin{tikzcd} p(\varphi,\theta,\tilde{\theta}) \arrow[r, "(d1)"'] \arrow[rr, bend left, "(d)"] & p_{smi,\eta}(\varphi,\theta,\tilde{\theta} \mid Y_1) \arrow[r, "(d2)"'] & p_{smi,\eta}(\varphi,\theta,\tilde{\theta} \mid Y) \end{tikzcd}$$

Update (d) is the same as defined above in Equation 10

Updates (d1) and (d2) are similar to Eq. 7, substituting the corresponding Y and current state of beliefs

$$
\begin{aligned}
p_{(d1)}(\varphi,\theta,\tilde{\theta}\mid Y_1) =& p(Y_1\mid\varphi,\tilde{\theta})^{\eta} p(Y_1\mid\varphi,\theta)\ \frac{1}{p(Y_{1,\eta})_{\tilde{\theta}}\ p(Y_1\mid\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) \\
\propto& p(Y_1\mid\varphi,\tilde{\theta})^{\eta} p(Y_1\mid\varphi,\theta)\ \frac{1}{p(Y_1\mid\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) \\
=& p(Y_1\mid\varphi,\tilde{\theta})^{\eta}\ p(Y_1\mid\varphi,\theta) \frac{p(\varphi)}{\int p(Y_1\mid\varphi,\theta)\ p(\varphi,\theta)\ d\theta} p(\varphi,\theta,\tilde{\theta}) \\
p_{(d1)+(d2)}(\varphi,\theta,\tilde{\theta}\mid Y_1,Y_2) \propto& p(Y_2\mid\varphi,\tilde{\theta})^{\eta}\ p(Y_2\mid\varphi,\theta) \frac{p_{(d1)}(\varphi)}{\int p(Y_2\mid\varphi,\theta)\ p_{(d1)}(\varphi,\theta)\ d\theta} p_{(d1)}(\varphi,\theta,\tilde{\theta}) \\
\propto& \left(p(Y_1\mid\varphi,\tilde{\theta})\ p(Y_2\mid\varphi,\tilde{\theta})\right)^{\eta} \left(p(Y_1\mid\varphi,\theta)\ p(Y_2\mid\varphi,\theta)\right)\cdot \\
&\cdot \frac{p_{(d1)}(\varphi)}{\int p(Y_2\mid\varphi,\theta)\ p_{(d1)}(\varphi,\theta)\ d\theta} \frac{1}{p(Y_1\mid\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) \\
\propto& p(Y\mid\varphi,\tilde{\theta})^{\eta} p(Y\mid\varphi,\theta)\ \frac{1}{p(Y\mid\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}).
\end{aligned}
$$

from here is clear that $p_{(d1)+(d2)}(\varphi,\theta,\tilde{\theta}\mid Y_1,Y_2) = p_{(d)}(\varphi,\theta,\tilde{\theta}\mid Y)$. In the last step, we used the following identity

$$
\begin{aligned}
\frac{p_{(d1)}(\varphi)}{\int p(Y_2\mid\varphi,\theta)\ p_{(d1)}(\varphi,\theta)\ d\theta} &= \frac{\int\int p_{(d1)}(\varphi,\theta,\tilde{\theta}) d\theta d\tilde{\theta}}{\int\int p(Y_2\mid\varphi,\theta)\ p_{(d1)}(\varphi,\theta,\tilde{\theta})\ d\theta d\tilde{\theta}} \\
&\propto \frac{\int\int p(Y_1\mid\varphi,\tilde{\theta})^{\eta}\ p(Y_1\mid\varphi,\theta) \frac{1}{p(Y_1\mid\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) d\theta d\tilde{\theta}}{\int\int p(Y_2\mid\varphi,\theta)\ p(Y_1\mid\varphi,\tilde{\theta})^{\eta}\ p(Y_1\mid\varphi,\theta) \frac{1}{p(Y_1\mid\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta})\ d\theta d\tilde{\theta}} \\
&= \frac{\frac{1}{p(\varphi)} \left(\int p(Y_1\mid\varphi,\tilde{\theta})^{\eta} p(\varphi,\tilde{\theta}) d\tilde{\theta}\right) \left(\int p(Y_1\mid\varphi,\theta) p(\varphi,\theta) d\theta\right)}{\frac{1}{p(\varphi)} \left(\int p(Y_1\mid\varphi,\tilde{\theta})^{\eta} p(\varphi,\tilde{\theta})\ d\tilde{\theta}\right) \left(\int p(Y_1\mid\varphi,\theta) p(Y_2\mid\varphi,\theta)\ p(\varphi,\theta)\ d\theta\right)} \\
&= \frac{p(Y_1\mid\varphi)_{\theta}}{p(Y\mid\varphi)_{\theta}}
\end{aligned}
$$

here again, we assumed that θ and $\tilde{\theta}$ are conditionally independent given φ in the prior, so $p(\varphi,\theta,\tilde{\theta}) = p(\varphi,\theta)\ p(\varphi,\tilde{\theta})\frac{1}{p(\varphi)}$.

4.2 Targeting the modular posterior

4.2.1 Detailed balance of SMI posterior

Here we show that the η-smi posterior (and cut posterior in particular) preserves the detailed balance condition when we use the transition kernel implied by the two-stage MCMC algorithm proposed in the main text.

1. Sample $(\varphi,\tilde{\theta}) \sim p_{pow,\eta}(\varphi,\tilde{\theta}\mid Z,Y) = p(Z\mid\varphi)\ p(Y\mid\varphi,\tilde{\theta})^{\eta} \frac{1}{p(Z,Y_{\eta})_{\tilde{\theta}}} p(\varphi,\tilde{\theta})$
2. Sample $\theta \sim p(\theta\mid Y,\varphi) = p(Y\mid\varphi,\theta) \frac{1}{p(Y\mid\varphi)_{\theta}} p(\theta)$

The first step updates $(\varphi,\tilde{\theta})$ using the powered likelihood. It is not difficult to target this posterior using traditional sampling methods.

The second term updates θ exactly from its conditional posterior given data Y from module 2, and a *fixed* value φ.

The transition kernel for one iteration in this scheme is given by

$$K(\varphi',\theta',\tilde{\theta}'\mid\varphi,\theta,\tilde{\theta}) = K(\varphi',\tilde{\theta}'\mid\varphi,\tilde{\theta}) p(\theta'\mid Y,\varphi'). \tag{11}$$

By construction, the first stage of the update (using the powered likelihood) is in detailed balance with the powered likelihood, i.e. satisfies

$$p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)K(\varphi', \tilde{\theta}' \mid \varphi, \tilde{\theta}) = p_{pow,\eta}(\varphi', \tilde{\theta}' \mid Z, Y)K(\varphi, \tilde{\theta} \mid \varphi', \tilde{\theta}')$$

From here we see that the η-smi posterior in Eq. 3 satisfies detailed balance with the transition kernel in eq. 11

$$\begin{aligned}
&p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)K(\varphi', \theta', \tilde{\theta}' \mid \varphi, \theta, \tilde{\theta}) \\
&\quad= [p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)p(\theta \mid Y, \varphi)][K(\varphi', \tilde{\theta}' \mid \varphi, \tilde{\theta})p(\theta' \mid Y, \varphi')] \\
&\quad= p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)p(\theta \mid Y, \varphi)\frac{p_{pow,\eta}(\varphi', \tilde{\theta}' \mid Z, Y)K(\varphi, \tilde{\theta} \mid \varphi', \tilde{\theta}')}{p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)}p(\theta' \mid Y, \varphi') \\
&\quad= p_{pow,\eta}(\varphi', \tilde{\theta}' \mid Z, Y)p(\theta' \mid Y, \varphi')K(\varphi, \tilde{\theta} \mid \varphi', \tilde{\theta}')p(\theta \mid Y, \varphi) \\
&\quad= p_{smi,\eta}(\varphi', \theta', \tilde{\theta}' \mid Z, Y)K(\varphi, \theta, \tilde{\theta} \mid \varphi', \theta', \tilde{\theta}')
\end{aligned}$$

4.2.2 Further details about Nested MCMC for SMI posterior.

Our implementation of MCMC targeting the SMI posterior is a two-stage sampler described in Algorithm 1 in the main article.

This is arguably the simpler approach that has been discussed in literature about MCMC targeting a modular posterior (e.g. Unbiased MCMC via couplings (Jacob et al., 2017)).

Algorithm 1 does not make specific assumptions about the class of MCMC sampler that is used at each stage. The only requirement is to check convergence in both stages. For the first step, sampling ϕ, we proceed as any traditional implementation of MCMC, running the chains until we have satisfied classical convergence test. The second step, sampling θ, we only need to guarantee that the last sample in every sub-chain are taken after we reached the equilibrium distribution.

Our examples in Section 5 use standard MCMC at each step. For the agricultural data in Sec.5.2 we use random walk Metropolis-Hastings in both stages. For the epidemiological data, the samplers for both stages are implemented in Stan (Carpenter et al., 2017), using Hamiltonian Monte Carlo. In both cases we performed convergence analysis for the main Chain (step 1), and experimented with various lengths for the sub-chain, in the end, we choose 500 iterations as a conservative length that guaranteed the last iteration was sampled from the equilibrium distribution. The detailed implementation can be found in the accompanying R package `aistats2020smi` available in GitHub [1].

4.4 Computational cost of SMI

Our baseline is standard Bayes-MCMC on the original full model. We suppose for simplicity this was implemented using separate updates for $\theta|\varphi$ and $\varphi|\theta$, though these need not be Gibbs updates. Let $\tau_{\varphi,\theta}$ be the Integrated Autocorrelation Time (IACT) of Bayes-MCMC. If the Effective Sample Size (ESS) of the full Bayes-MCMC output is N then we must have done $T = N\tau_{\varphi,\theta}$ MCMC steps. If *one Bayes-MCMC step updating both θ and φ has unit cost* then the overall cost is $W_{bm} = T$[time]. This doesn't parallelise.

The SMI posterior is

$$p_{\eta-smi}(\varphi, \tilde{\theta}, \theta|Y, Z) = p_{pow,\eta}(\varphi, \tilde{\theta}|Y, Z)p(\theta|Y, \varphi).$$

If we use the same Bayes-MCMC updates to sample $p_{pow,\eta}(\varphi, \tilde{\theta}|Y, Z)$ then the work sampling $(\varphi, \tilde{\theta})$ at one η-value is W_{Bayes}. Thin the φ samples every $\tau_{\varphi,\theta}$ steps to get an ESS about N (this is rough, because the target changes with η, but reasonable if we allow for some tuning of the MCMC with η - we didnt need to tune in our examples).

We use the Bayes-MCMC update for $p(\theta|Y, \varphi)$ in SMI. Let τ_θ be the IACT. Typically, $\tau_\theta < \tau_{\varphi,\theta}$ (the target has lower dimension; illustrative proofs can be given in simple special cases) so take $\tau_\theta = \tau_{\varphi,\theta}$ (conservative, and note that these "side-chains" targeting $p(\theta|Y, \varphi)$ can be started close to equilibrium using the θ-value output from sampling at the previous step). We run the θ-sampler to equilibrium (initialise $\theta^{(t)}$-run with $\theta^{(t-1)}$). Suppose this takes $K\tau_\theta$ steps ($K \approx 5$ is reasonable).

[1] `https://github.com/christianu7/aistats2020smi`

In the following we assume simulation at different η-values is parallelized over machines, while simulation of $\theta|\phi$ is parallelized over threads on a machine. Suppose we have M_t threads on each of M_p machines. The θ-sampling parallelizes with a small communication overhead if the time to do $K\tau_{\varphi,\theta}$ of the $\theta|\varphi$-updates is significantly larger than the communication time. The cost of the θ-update in the second stage of SMI is no more than the cost of one update in the original Bayes-MCMC where both θ and φ were updated, so a runtime cost for θ-updates equal one unit is conservative. It follows that $(\tilde{\theta}, \varphi, \theta)$-sampling at one η-value costs about

$$W_\eta = W_{bm}(1 + K/M_t).$$

We have to repeat this J times, sampling $p_{\eta-smi}(\varphi, \tilde{\theta}, \theta|Y, Z)$ for each of J different η-values spaced between $\eta = 0$ and $\eta = 1$ ($J \approx 20$ should be enough) using M_p machines. This larger task parallelises essentially perfectly. The total cost is

$$W_{smi} = W_{bm}(1 + K/M_t) \times J/M_p.$$

For eg if we assign resources as $M_p = J$ and $M_t = 1$ (just parallelise over η) the SMI cost is not worse than about 10 times the cost of doing Bayes-MCMC. This reflects our experience.

Finally, we compute and smooth the WAIC across the J runs at different η-values. This part is fast output-analysis. The ESS must be big enough to get stable WAIC estimates, but WAIC is "nice" to estimate. Very roughly,

$$W_{smi} \simeq 10 W_{bm}$$

should be achievable without a great deal of work on top of the cost of implementing and running standard Bayes-MCMC.

5 Data Analyses

5.1 Simulation study: Biased data

Model:

$$Z \mid \varphi \sim N(\varphi, \sigma_z^2)$$
$$Y \mid \varphi, \theta \sim N(\varphi + \theta, \sigma_y^2)$$

with σ_z^2 and σ_y^2 (and other σ's) known.

Priors:

$$\varphi \sim N(0, \sigma_\varphi^2)$$
$$\theta \sim N(0, \sigma_\theta^2)$$
$$\tilde{\theta} \sim N(0, \tilde{\sigma}_\theta^2)$$

In our example on the main text, we know the generative parameters: $(\varphi^*, \theta^*, \tilde{\theta}^*)$. We compare these *true* values with the estimates arising from the η-smi posterior, for different values of $\eta \in [0, 1]$.

5.1.1 SMI posterior

First, derive $p(Y \mid \varphi)_\theta$ as a function of φ

$$\begin{aligned} p(Y \mid \varphi)_\theta &= \frac{1}{p(\varphi)} \int p(Y \mid \varphi, \theta)\, p(\varphi, \theta)\, d\theta \\ &\propto \exp\{-\frac{1}{2}[\varphi^2(\frac{m}{m\sigma_\theta^2 + \sigma_y^2}) - 2\varphi(\bar{Y}\frac{m}{m\sigma_\theta^2 + \sigma_y^2})]\} \end{aligned}$$

Now we can obtain the η-smi posterior

$$\begin{aligned} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) &= p(Z \mid \varphi)\, p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta)\, \frac{1}{p(Z, Y_\eta)_{\tilde{\theta}}\, p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}) \\ &\propto \exp\{-\frac{1}{2}\, [\varphi^2(\frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) + \frac{1}{\sigma_\varphi^2} - \frac{m}{\sigma_y^2 + m\sigma_\theta^2}) - 2\varphi(\frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) - \bar{Y}\frac{m}{\sigma_y^2 + m\sigma_\theta^2}) + \\ &\qquad \theta^2(\frac{m}{\sigma_y^2} + \frac{1}{\sigma_\theta^2}) - 2\theta(\bar{Y}\frac{m}{\sigma_y^2}) + \\ &\qquad \tilde{\theta}^2(\eta\frac{m}{\sigma_y^2} + \frac{1}{\tilde{\sigma}_\theta^2}) - 2\tilde{\theta}(\eta\bar{Y}\frac{m}{\sigma_y^2}) + \\ &\qquad + 2\varphi\theta(\frac{m}{\sigma_y^2}) + 2\varphi\tilde{\theta}(\eta\frac{m}{\sigma_y^2})]\} \end{aligned}$$

From here we see that the joint posterior distribution for $(\varphi, \theta, \tilde{\theta})$ is a multivariate normal distribution defined as:

$$p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) = \text{Normal}(\mu, \Sigma), \tag{12}$$

with

$$\Sigma = \begin{bmatrix} \frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) - \frac{m}{\sigma_y^2 + m\sigma_\theta^2} + \frac{1}{\sigma_\varphi^2} & \frac{m}{\sigma_y^2} & \eta\frac{m}{\sigma_y^2} \\ \frac{m}{\sigma_y^2} & \frac{m}{\sigma_y^2} + \frac{1}{\sigma_\theta^2} & 0 \\ \eta\frac{m}{\sigma_y^2} & 0 & \eta\frac{m}{\sigma_y^2} + \frac{1}{\tilde{\sigma}_\theta^2} \end{bmatrix}^{-1}, \text{ and } \mu = \Sigma \begin{bmatrix} \frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) - \bar{Y}\frac{m}{\sigma_y^2 + m\sigma_\theta^2} \\ \bar{Y}\frac{m}{\sigma_y^2} \\ \eta\bar{Y}\frac{m}{\sigma_y^2} \end{bmatrix}.$$

The generative parameters described in the main text are as follows

```
> n=25 # Sample size for Z
> m=50 # Sample size for Y
> phi = 0
> theta = 1 # bias
> sigma_z = 2 # variance for Z
> sigma_y = 1 # variance for Y
```

The true bias is $\theta = 1$. Assume we have an over-optimistic view of the bias, with prior distribution centered in 0 and relatively small prior variance.

In Figure 2 we show posterior distributions (mean $\pm$ std. dev.) for a randomly generated dataset ($\bar{Z}$ =-0.0667; $\bar{Y}$ =1.0562) using the generative parameters described in the main text. Note that the conventional bayes ($\eta = 1$) is the worst estimation for the true parameters φ accross all posible candidates $\eta \in [0, 1]$.

```
> # Posterior for conventional bayes eta=1
> posterior = SMI_post_biased_data( Z=Z, Y=Y,
+                                   sigma_z=sigma_z, sigma_y=sigma_y,
+                                   sigma_phi=sigma_phi,
+                                   sigma_theta=sigma_theta, sigma_theta_tilde=sigma_theta,
+                                   eta=1 )
> posterior = mapply('rownames<-', posterior, MoreArgs=list(value=param_names))
> # posterior mean
> posterior$mean

                 [,1]
phi         0.3511440
theta       0.6528197
theta_tilde 0.6528197
```

5.1.2 Mean Square Error (MSE)

From Equation 3, we can compute the Posterior Squared Error (SE) of estimates arising from the SMI posterior φ, θ and , $\tilde{\theta}$,

$$\begin{aligned} SE(\varphi) &= \Sigma_{[1,1]} + (\mu_{[1]} - \varphi^*)^2 \\ SE(\theta) &= \Sigma_{[2,2]} + (\mu_{[2]} - \theta^*)^2 \\ SE(\tilde{\theta}) &= \Sigma_{[3,3]} + (\mu_{[3]} - \theta^*)^2 \end{aligned}$$

In the first simulation study of the main text we display the Mean Squared Error (MSE) for φ and θ, which is the result of averaging the posterior SE across datasets. We show that we can reach smaller MSE with values of η other than 0 and 1. To generate these plots, we produced 1000 synthetic datasets, computed MSE using Eq. 3 on each one, with a grid of values of $\eta \in [0, 1]$. The MSE lines displayed correspond to the *average* MSE across datasets, for each value of η.

Here we also show a comparison between θ and $\tilde{\theta}$. Figure 4 shows that $\tilde{\theta}$ is dominated by θ in MSE. The comparison emphasises the convenience of SMI over power likelihood discussed in section 4.4 of the main text.

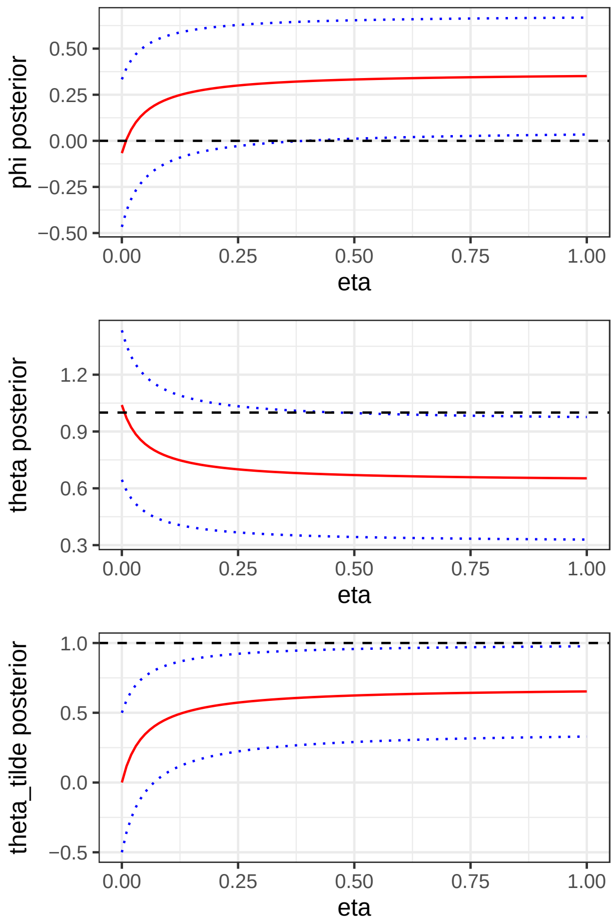

Figure 2: Posterior distribution of φ, θ and $\tilde{\theta}$ for a single dataset $(Z{,}Y)$. A black horizontal line shows the true generative value. The posterior mean is the solid red line and we show intervals with $\pm$ one posterior std. dev. using dotted blue lines.

5.1.3 Expected log pointwise predictive density (elpd)

The elpd is

$$elpd = \int \int p^*(z, y) \log p_{smi,\eta}(z, y \mid Z, Y) dzdy$$

where p^* is the distribution representing the true data-generating process and

$$p_{smi,\eta}(z, y \mid Z, Y) = \int \int p(z, y \mid \varphi, \theta)\, p_{smi,\eta}(\varphi, \theta \mid Y, Z)\, d\varphi\, d\theta$$

is a candidate posterior predictive distribution, indexed by η.

Let $\begin{bmatrix} a & b \\ b & c \end{bmatrix} = Cov(\varphi, \theta \mid Z, Y)^{-1}$ be the inverse of the posterior covariance matrix of (φ, θ), and $\begin{bmatrix} d \\ e \end{bmatrix} = E(\varphi, \theta \mid Z, Y)$ the posterior means.

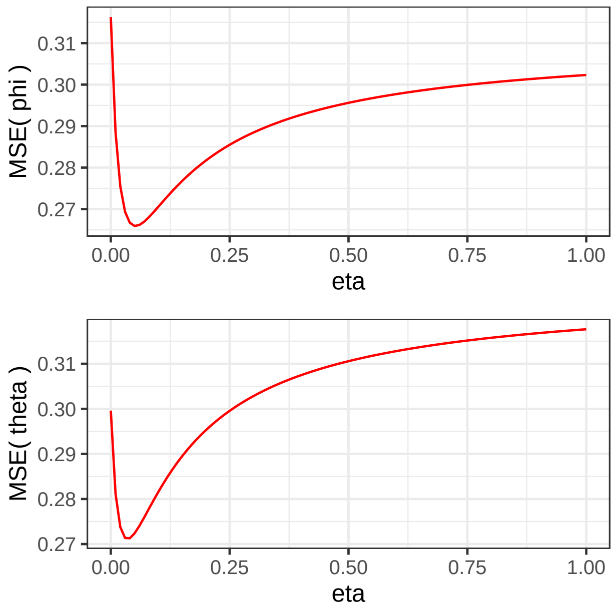

Figure 3: Mean Squared Error of the two main parameters under SMI posterior.

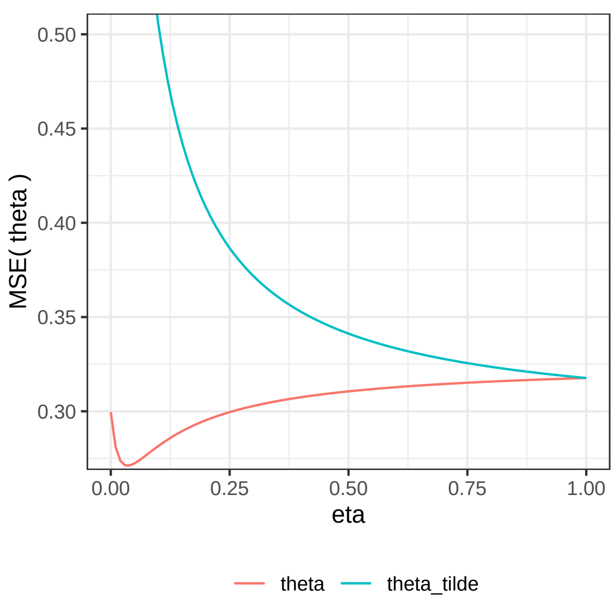

Figure 4: Comparison of the bias estimation under SMI (theta) vs powered likelihood (theta tilde)

Following straightforward Gaussian completion we can show that the joint posterior distribution for φ, θ, and new data z_0 and y_0 is:

$$
\begin{aligned}
p_{smi,\eta}(z_0, y_0, \varphi, \theta|Z, Y) &\propto p(z_0, y_0|\varphi, \theta)\ p_{smi,\eta}(\varphi, \theta|Z, Y) \\
&\propto \exp\{-\frac{1}{2}\ [z_0^2(\frac{1}{\sigma_z^2}) + y_0^2(\frac{1}{\sigma_y^2}) + \varphi^2(a + \frac{1}{\sigma_z^2} + \frac{1}{\sigma_y^2}) + \theta^2(c + \frac{1}{\sigma_y^2}) + \\
&\qquad - 2z_0\varphi(\frac{1}{\sigma_z^2}) - 2y_0\varphi(\frac{1}{\sigma_y^2}) - 2y_0\theta(\frac{1}{\sigma_y^2}) + 2\varphi\theta(b + \frac{1}{\sigma_y^2}) \\
&\qquad - 2\varphi(ad + be) - 2\theta(bd + ce)]\}
\end{aligned}
$$

So we have

$$p_{smi,\eta}(z_0, y_0, \varphi, \theta \mid Z, Y) = \text{Normal}(\mu, \Sigma), \tag{13}$$

with

$$\Sigma = \begin{bmatrix} \frac{1}{\sigma_z^2} & 0 & -\frac{1}{\sigma_z^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} & -\frac{1}{\sigma_y^2} & -\frac{1}{\sigma_y^2} \\ -\frac{1}{\sigma_z^2} & -\frac{1}{\sigma_y^2} & a + \frac{1}{\sigma_z^2} + \frac{1}{\sigma_y^2} & b + \frac{1}{\sigma_y^2} \\ 0 & -\frac{1}{\sigma_y^2} & b + \frac{1}{\sigma_y^2} & c + \frac{1}{\sigma_y^2} \end{bmatrix}^{-1}, \text{ and } \mu = \Sigma \begin{bmatrix} 0 \\ 0 \\ ad + be \\ bd + ce \end{bmatrix}.$$

We know the true generative values φ^* and θ^*, so we can compute *elpd* using Monte Carlo samples from the true generative distribution p^* and evaluate this values in the log-density of the bivariate normal $(z_0, y_0|Z, Y)$ from Equation 13.

In Figure 5 we show the Monte Carlo estimation of the *elpd*. We select the optimal η as the value that maximise the *elpd*. To generate this plot, we produced 1000 synthetic datasets, computed *elpd* on each one (using Monte Carlo), with a grid of values of $\eta \in [0, 1]$. The *elpd* line correspond to the *average* elpd across datasets, for each value of η.

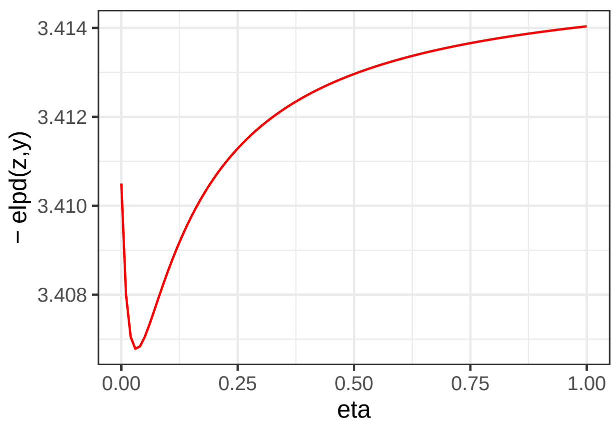

Figure 5: ELPD under SMI posterior.

5.2 Agricultural data

The aim of the study, described in detail in Styring et al. (2017), is to provide statistical evidence about a specific agricultural practice of the first urban centres in northern Mesopotamia.

The hypothesis is that increased agricultural production to support growing urban populations was achieved by cultivation of larger areas of land, entailing lower manure/midden inputs per unit area. This practice is known as *extensification*.

Our contribution goes into extending the methods used to perform Bayesian analysis in this adverse scenario of model misspecification and big missing data.

5.2.1 Data

The data consists of measurements of nitrogen and carbon isotopes for a collection of crop remains. There are two datasets: archaeological and modern, which we will denote by $\mathcal{A}$ and $\mathcal{M}$, respectively. First, the *Archaeological* dataset $\mathcal{A}$, consists of data gathered from excavations of antique crop sites in the region of Mesopotamia. Second, the *Modern* dataset, $\mathcal{M}$, was gathered in a controlled experimental setting in recent years. Further characteristics and description of variables on each dataset can be found in the statistical supplement to Styring et al. (2017).

5.2.2 The model

We preserved the model stated in Styring et al. (2017). See Figure 6. The main goal of that study is to test the hypothesis of *extensification* in the ancient urban sites. This hypothesis can be condensed in analysing the strength of the effect of the site size S_i on the corresponding manure level M_i for the records in the archaeological data, $\{i \in \mathcal{A}\}$. The more negative the estimated effect, the stronger the evidence supporting the extensification hypothesis.

However, there is no available information about the Manuring levels in the ancient dataset. This is addressed by using a *Data augmentation* perspective, and consider that all the corresponding values of the manure level in the Archaeological data are missing.

The model consists of a two-module model, which integrates archaeological and calibration data so the missing manuring levels can be inferred. The first module is a Proportional Odds model (PO), with the missing Manure Levels in the archaeological data as the ordinal response ($M_i; i \in \mathcal{A}$). The second module consists of a linear Gaussian model (HM), applicable to both datasets, with the Nitrogen level of the crops (Z) as the response, and Manure levels as one of the predcitors. The graphical representation of the model is depicted in Figure 6

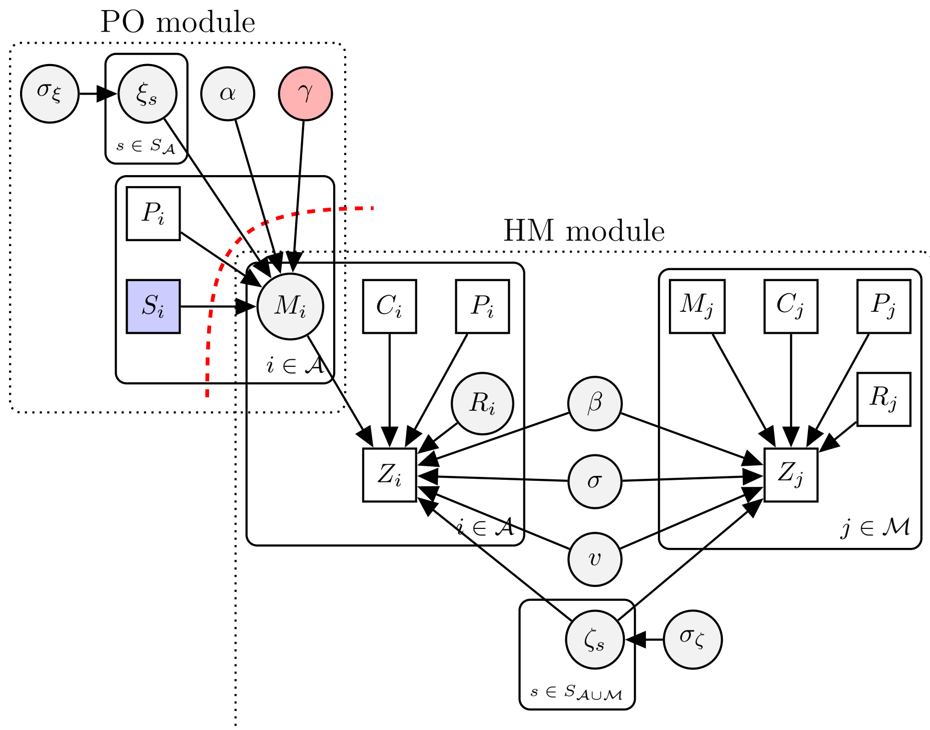

Figure 6: Graphical representation of the model for the agricultural data. Squares denote observable variables and circles denote unknown quantities (parameters and missing data).The main interest of the study is on the parameter γ (red circle), effect of size S_i (blue square) on Manure level M_i. The dashed line indicates the cut where SMI is applied for the imputation of missing manure.

The model can be studied from a multi-modular perspective. Indeed, the supplement of Styring et al. (2017) performs Bayesian Multiple Imputation (BMI) to impute the missing manure levels with parameters learnt from the HM module, and therefore *cutting* the influence from the PO module into this imputation. The parameters of the PO module are then inferred conditional on the imputed values and other information in the archaeological data.

In Figure 7 we draw a simplified version of the complete model to clarify how the SMI framework can be applied in this setting. The mapping of variables between the complete and simplified graphs is as follows: the missing manure levels will take the role of ϕ; the observed data in the HM module is Z, the parameters in the PO module will be θ, and the rest of archaeological data in the PO module is Y.

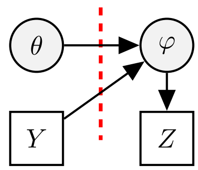

Figure 7: Simplified representation of the model for the agricultural data.

The simplified version resembles our graphical model in Figure 1. From here it is clear how the Bayesian imputation approach taken by Styring et al. (2017) is equivalent to a cut model: first learn ϕ/manure-level from the HM module, and then learning θ (ie γ) conditional on ϕ and Y. This scheme yields our known cut posterior from Eq. 2.

$$P_{cut}(\varphi, \theta \mid Z, Y) = P(\varphi \mid Z)\ P(\theta \mid Y, \varphi)$$

We extend the Bayesian imputation approach in Styring et al. (2017) and apply Semi-Modular Inference to their setting.

$$P_{smi,\eta}(\varphi, \theta \mid Z, Y) = \int P_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)\ P(\theta \mid Y, \varphi) d\tilde{\theta}$$

SMI allows us to expand the space of candidate posteriors in a way that we can control smoothly, rather that eliminating the influence of the PO module in the imputation of the missing Manure Levels.

`gamma_po_1`

In Figure 8 we display the collection of candidate distributions spanned by SMI by considering a grid of values for $\eta \in [0, 1]$. The distribution at η =0 is comparable to the Bayesian Imputation approach in Styring et al. (2017).

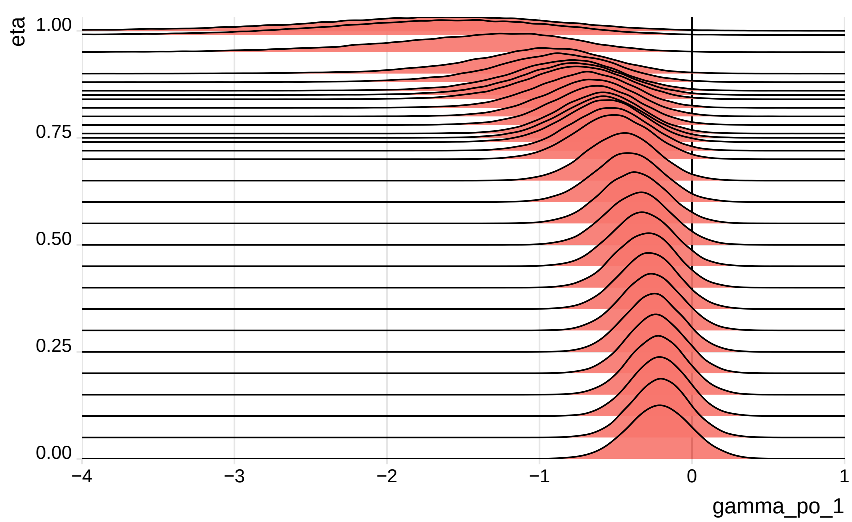

Figure 8: SMI posterior distribution for the parameter of interest in the agricultural model.

To choose the optimal value $\eta = \eta^*$ say, we maximise the Expected Log Predictive Density or *ELPD* (Vehtari et al., 2017) to predict a new value of the response in the HM module ($Z_i; i \in \mathcal{A}$ in the diagram). In Figure 9 we show the negative ELPD as a function of η. The optimal value is reached at η^* =0.82.

The ELDP changes dramatically over the range $0 \leq \eta \leq 1$ and η^* is clearly distinguished from 0 or 1. This impacts downstream inference: the Bayes factor of interest changes from $BF_{smi,0}(\gamma \leq 0)$ =5.54 in the cut model, to $BF_{smi,0.82}(\gamma \leq 0)$ =117.11 in the optimal SMI posterior with η =0.82, a substantial shift in the strength of the evidence for the *extensification* hypothesis. In Figure 10 we show the comparison of Bayes Factors across values of $\eta \in [0, 1]$ for the hypothesis $\gamma \leq 0$.

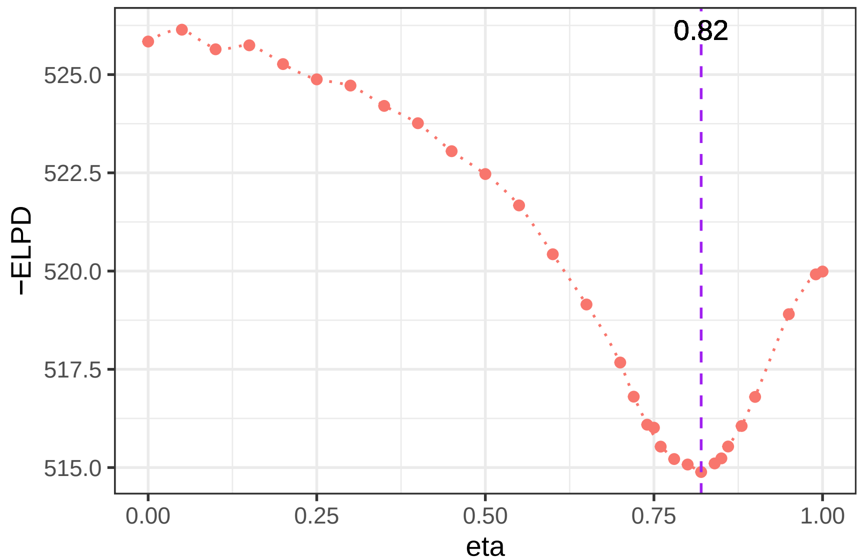

Figure 9: Choosing the best SMI posterior candidate by choosing the η-value maximizing the ELPD.

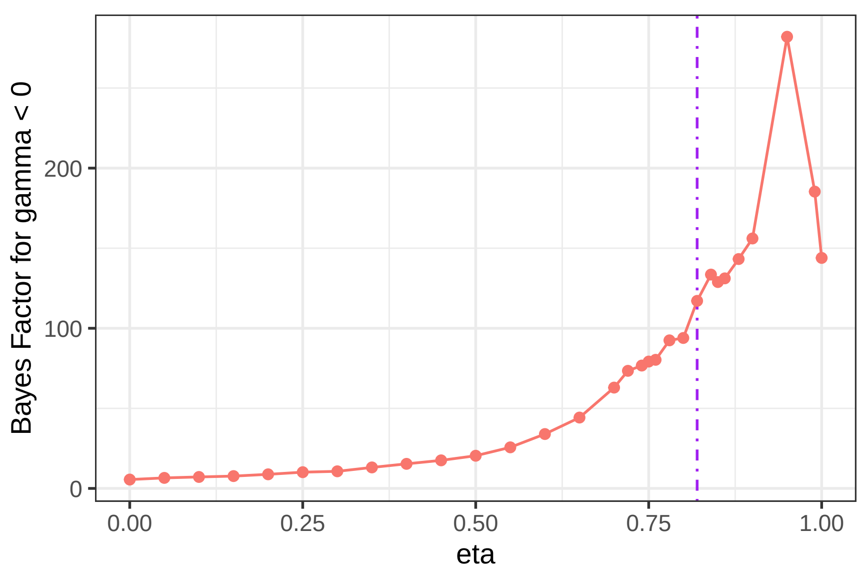

Figure 10: Bayes Factor for the hypothesis $\gamma \leq 0$ across all values of η.

References

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, 76(1):1–32.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2017). Unbiased Markov chain Monte Carlo with couplings.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1):37–43.

Styring, A. K., Charles, M., Fantone, F., Hald, M. M., McMahon, A., Meadow, R. H., Nicholls, G. K., Patel, A. K., Pitre, M. C., Smith, A., So?tysiak, A., Stein, G., Weber, J. A., Weiss, H., and Bogaard, A. (2017). Isotope evidence for agricultural extensification reveals how the world's first cities were fed. *Nature Plants*, 3(6).

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.