Better Together? Statistical Learning in Models Made of Modules

Pierre E. Jacob, Lawrence M. Murray, Chris C. Holmes and Christian P. Robert

Discussed by: Chris. J. Oates
Newcastle University
Lloyds-Turing Programme on Data-Centric Engineering

December 2017 @ PNWG
Background
According to Cockayne et al. (2017), a probabilistic numerical method (PNM) consists of:

- an (often $\infty$-dimensional) state space $\mathcal{X}$
- a (finite dimensional) information space $\mathcal{A}$
- a quantity of interest space $\mathcal{Q}$
- an information operator $A: \mathcal{X} \to \mathcal{A}$
- a belief update operator $B: \mathcal{P}_{\mathcal{X}} \times \mathcal{A} \to \mathcal{P}_{\mathcal{Q}}$

This general set-up is somewhat similar to Bissiri, Holmes and Walker (2016). Also similar to an old framework; Zellner (1988) calls $B$ an information processing rule (IPR).


Cockayne et al. (2017) call a PNM Bayesian if

$$B(\mu, a) = Q_{#}\mu^a$$

Key motivation for Bayesian PNM is coherence and interpretability over pipelines - can be argued in detail.
According to Cockayne et al. (2017), a probabilistic numerical method (PNM) consists of:

- an (often $\infty$-dimensional) state space $\mathcal{X}$
- a (finite dimensional) information space $\mathcal{A}$
- a quantity of interest space $\mathcal{Q}$
- an information operator $A : \mathcal{X} \to \mathcal{A}$
- a belief update operator $B : \mathcal{P}_\mathcal{X} \times \mathcal{A} \to \mathcal{P}_\mathcal{Q}$

This general set-up is somewhat similar to Bissiri, Holmes and Walker (2016). Also similar to an old framework; Zellner (1988) calls $B$ an information processing rule (IPR).


Cockayne et al. (2017) call a PNM Bayesian if

$$B(\mu, a) = Q \# \mu^a$$

Key motivation for Bayesian PNM is coherence and interpretability over pipelines - can be argued in detail.
According to Cockayne et al. (2017), a probabilistic numerical method (PNM) consists of:

- an (often \(\infty\)-dimensional) state space \(\mathcal{X}\)
- a (finite dimensional) information space \(\mathcal{A}\)
- a quantity of interest space \(\mathcal{Q}\)
- an information operator \(A : \mathcal{X} \rightarrow \mathcal{A}\)
- a belief update operator \(B : \mathcal{P}_\mathcal{X} \times \mathcal{A} \rightarrow \mathcal{P}_\mathcal{Q}\)

This general set-up is somewhat similar to Bissiri, Holmes and Walker (2016). Also similar to an old framework; Zellner (1988) calls \(B\) an information processing rule (IPR).


Cockayne et al. (2017) call a PNM Bayesian if

\[ B(\mu, a) = Q # \mu^a \]

Key motivation for Bayesian PNM is coherence and interpretability over pipelines - can be argued in detail.
Moreover Zellner (1988) argues that the Bayesian update is optimal:

*Some IPR’s may be in efficient in the sense that the output information, measured in a suitable metric, is less than the input information. On the other hand, some IPR’s may add extraneous information so that the output information is greater than the given input information, an undesirable state of affairs.*

In equations, Zellner defined an optimal IPR as one that satisfies:

\[
\mathbb{H}[B(\mu, a)] + \mathbb{H}[A_{\#\mu}(a)] - \mathbb{H}[x \mapsto p(a|x)] - \mathbb{H}[\mu(x)] = 0
\]

However, the definition of Bayesian PNM is rather restrictive...
Moreover Zellner (1988) argues that the Bayesian update is optimal:

_Some IPR's may be in efficient in the sense that the output information, measured in a suitable metric, is less than the input information. On the other hand, some IPR's may add extraneous information so that the output information is greater than the given input information, an undesirable state of affairs._

In equations, Zellner defined an optimal IPR as one that satisfies:

\[
\mathbb{H}[B(\mu, a)] + \mathbb{H}[A#\mu(a)] - \mathbb{H}[x \mapsto p(a|x)] - \mathbb{H}[\mu(x)] = 0
\]

However, the definition of Bayesian PNM is rather restrictive...
Moreover Zellner (1988) argues that the Bayesian update is optimal:

*Some IPR's may be in efficient in the sense that the output information, measured in a suitable metric, is less than the input information. On the other hand, some IPR's may add extraneous information so that the output information is greater than the given input information, an undesirable state of affairs.*

In equations, Zellner defined an optimal IPR as one that satisfies:

\[
\mathbb{H}[B(\mu, a)] + \mathbb{H}[A#\mu(a)] - \mathbb{H}[x \mapsto p(a|x)] - \mathbb{H}[\mu(x)] = 0
\]

However, the definition of Bayesian PNM is rather restrictive...
Numerical Integration Thought Experiment

\[ \int_0^1 f(x) \, dx \quad \text{modelled as a r.v.} \]

\[ = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx \]

\[ \text{to computer 1} \quad \text{to computer 2} \]

\[ \downarrow \quad \downarrow \]

posterior 1 \quad posterior 2

\[ \downarrow \quad \downarrow \]

posterior \quad ? = \quad \text{posterior}

Q: When will the distributed version (RHS) return the same answer as the full version (LHS)?

For bigger computational work-flows the constraints are even more restrictive, and work against efficient computational architectures.

OTOH, splitting provides robustness to model misspecification - i.e. to using the wrong kernel. So potentially being non-Bayesian confers robustness to model misspecification.
Numerical Integration Thought Experiment

\[ \int_0^1 f(x) \, dx = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx \]

modelled as a r.v. to computer 1 to computer 2

\[ \text{posterior 1} \quad \downarrow \quad \text{posterior 2} \quad \downarrow \quad \text{posterior} \quad ? \quad \text{posterior} \]

Q: When will the distributed version (RHS) return the same answer as the full version (LHS)?

For bigger computational work-flows the constraints are even more restrictive, and work against efficient computational architectures.

OTOH, splitting provides robustness to model misspecification - i.e. to using the wrong kernel. So potentially being non-Bayesian confers robustness to model misspecification.
Numerical Integration Thought Experiment

\[ \int_0^1 f(x) \, dx = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx \]

modelled as a r.v.

\[ \rightarrow \quad \text{to computer 1} \quad \downarrow \quad \text{to computer 2} \]

\[ \downarrow \quad \text{posterior 1} \quad \downarrow \quad \text{posterior 2} \]

\[ \downarrow \quad ? \quad \downarrow \quad \text{posterior} \]

Q: When will the distributed version (RHS) return the same answer as the full version (LHS)?

For bigger computational work-flows the constraints are even more restrictive, and work against efficient computational architectures.

OTOH, splitting provides robustness to model misspecification - i.e. to using the wrong kernel. So potentially being non-Bayesian confers robustness to model misspecification.
Q: When will the distributed version (RHS) return the same answer as the full version (LHS)?

For bigger computational work-flows the constraints are even more restrictive, and work against efficient computational architectures.

OTOH, splitting provides robustness to model misspecification - i.e. to using the wrong kernel. So potentially being non-Bayesian confers robustness to model misspecification.
Numerical Integration Thought Experiment

\[ \int_0^1 f(x) \, dx = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx \]

modelled as a r.v.

to computer 1
to computer 2

\[ \downarrow \quad \downarrow \]

posterior 1
posterior 2

\[ \downarrow \quad \downarrow \]

posterior

Q: When will the distributed version (RHS) return the same answer as the full version (LHS)?

For bigger computational work-flows the constraints are even more restrictive, and work against efficient computational architectures.

OTOH, splitting provides robustness to model misspecification - i.e. to using the wrong kernel. So potentially being non-Bayesian confers robustness to model misspecification.
Numerical Integration Thought Experiment

\[ \int_0^1 f(x) \, dx = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx \]

modelled as a r.v. to computer 1 to computer 2

\[ \Downarrow \quad \Downarrow \]

posterior 1 posterior 2

\[ \Downarrow \quad \Downarrow \]

posterior = posterior

Q: When will the distributed version (RHS) return the same answer as the full version (LHS)?

For bigger computational work-flows the constraints are even more restrictive, and work against efficient computational architectures.

OTOH, splitting provides robustness to model misspecification - i.e. to using the wrong kernel. So potentially being non-Bayesian confers robustness to model misspecification.
**Numerical Integration Thought Experiment**

\[
\int_0^1 f(x) \, dx = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx
\]

modelled as a r.v. to computer 1 to computer 2

\[\Downarrow\]

posterior 1 posterior 2

\[\Downarrow\]

posterior \[?\] = posterior

Q: When will the distributed version (RHS) return the same answer as the full version (LHS)?

For bigger computational work-flows the constraints are even more restrictive, and work against efficient computational architectures.

OTOH, splitting provides robustness to model misspecification - i.e. to using the wrong kernel. So potentially being non-Bayesian confers robustness to model misspecification.
Less Trivial Thought Experiment

Partial differential equation with known coefficient $\theta$:

$$\nabla \cdot (\theta \nabla u) = f \quad \text{in } \Omega$$
$$u = 0 \quad \text{on } \partial \Omega$$

Standard Bayesian framework for estimation of $\theta$ based on a prior $p_\theta$ and noisy measurements:

$$y_i|\theta \sim \mathcal{N}(u(x_i; \theta), \sigma^2) \quad \text{i.i.d.}$$

Cockayne et al. (2016) considered MCMC over $\theta \in \Theta$. At each iteration of the Markov chain, the proposed parameter $\theta$ was related to data through use of a Bayesian PNM:

$$y_i|\theta \sim \int \mathcal{N}(u(x_i; \theta), \sigma^2) \mathcal{B}(\mu, \{f(x_1), \ldots, f(x_n)\})(du)$$

Cockayne et al. (2016) treated the forward problems as statistically independent. But that’s not really true, as MCMC recurrent.

OTOH, would we trust ourselves to elicit a prior for how the solution of the PDE depends on $\theta$? High risk of model misspecification if we tried that.
Less Trivial Thought Experiment

Partial differential equation with known coefficient $\theta$:

$$\nabla \cdot (\theta \nabla u) = f \text{ in } \Omega$$

$$u = 0 \text{ on } \partial \Omega$$

Standard Bayesian framework for estimation of $\theta$ based on a prior $p_\theta$ and noisy measurements:

$$y_i|\theta \sim \mathcal{N}(u(x_i; \theta), \sigma^2) \text{ i.i.d.}$$

Cockayne et al. (2016) considered MCMC over $\theta \in \Theta$. At each iteration of the Markov chain, the proposed parameter $\theta$ was related to data through use of a Bayesian PNM:

$$y_i|\theta \sim \int \mathcal{N}(u(x_i; \theta), \sigma^2) B(\mu, [f(x_1), \ldots, f(x_n)])(du)$$

Cockayne et al. (2016) treated the forward problems as statistically independent. But that’s not really true, as MCMC recurrent.

OTOH, would we trust ourselves to elicit a prior for how the solution of the PDE depends on $\theta$? High risk of model misspecification if we tried that.
Partial differential equation with known coefficient $\theta$:

$$\nabla \cdot (\theta \nabla u) = f \quad \text{in } \Omega$$

$$u = 0 \quad \text{on } \partial \Omega$$

Standard Bayesian framework for estimation of $\theta$ based on a prior $p_\theta$ and noisy measurements:

$$y_i|\theta \sim N(u(x_i; \theta), \sigma^2) \quad \text{i.i.d.}$$

Cockayne et al. (2016) considered MCMC over $\theta \in \Theta$. At each iteration of the Markov chain, the proposed parameter $\theta$ was related to data through use of a Bayesian PNM:

$$y_i|\theta \sim \int N(u(x_i; \theta), \sigma^2) B\left(\mu, [f(x_1), \ldots, f(x_n)]\right)(du)$$

Cockayne et al. (2016) treated the forward problems as statistically independent. But that’s not really true, as MCMC recurrent.

OTOH, would we trust ourselves to elicit a prior for how the solution of the PDE depends on $\theta$? High risk of model misspecification if we tried that.
Partial differential equation with known coefficient $\theta$:

$$\nabla \cdot (\theta \nabla u) = f \quad \text{in } \Omega$$

$$u = 0 \quad \text{on } \partial \Omega$$

Standard Bayesian framework for estimation of $\theta$ based on a prior $p_\theta$ and noisy measurements:

$$y_i | \theta \sim \mathcal{N}(u(x_i; \theta), \sigma^2) \quad \text{i.i.d.}$$

Cockayne et al. (2016) considered MCMC over $\theta \in \Theta$. At each iteration of the Markov chain, the proposed parameter $\theta$ was related to data through use of a Bayesian PNM:

$$y_i | \theta \sim \int \mathcal{N}(u(x_i; \theta), \sigma^2) B(\mu, [f(x_1), \ldots, f(x_n)])(du)$$

Cockayne et al. (2016) treated the forward problems as statistically independent. But that’s not really true, as MCMC recurrent.

OTOH, would we trust ourselves to elicit a prior for how the solution of the PDE depends on $\theta$? High risk of model misspecification if we tried that.
Less Trivial Thought Experiment

Partial differential equation with known coefficient $\theta$:

$$\nabla \cdot (\theta \nabla u) = f \quad \text{in } \Omega$$

$$u = 0 \quad \text{on } \partial \Omega$$

Standard Bayesian framework for estimation of $\theta$ based on a prior $p_\theta$ and noisy measurements:

$$y_i|\theta \sim \mathcal{N}(u(x_i; \theta), \sigma^2) \quad \text{i.i.d.}$$

Cockayne et al. (2016) considered MCMC over $\theta \in \Theta$. At each iteration of the Markov chain, the proposed parameter $\theta$ was related to data through use of a Bayesian PNM:

$$y_i|\theta \sim \int \mathcal{N}(u(x_i; \theta), \sigma^2) \mathcal{B} \left( \mu \big[ f(x_1), \ldots, f(x_n) \big] \right)(d\theta)$$

Cockayne et al. (2016) treated the forward problems as statistically independent. But that’s not really true, as MCMC recurrent.

OTOH, would we trust ourselves to elicit a prior for how the solution of the PDE depends on $\theta$? High risk of model misspecification if we tried that.
Is Bayesian PNM the unattainable dream?

fully Bayesian PNM pipeline in Oates, Cockayne and Aykroyd (2017)
vector-valued quadrature work of FX (in prep.)

Tübingen & co. pursuing a Gaussian message propagation - much more practical.

Or, can we justify a relaxation of Bayesian PNM? “Cutting some of the dependencies”?

Turns out this is not a new problem at all...
A Conundrum

Is Bayesian PNM the unattainable dream?

fully Bayesian PNM pipeline in Oates, Cockayne and Aykroyd (2017)
vector-valued quadrature work of FX (in prep.)

Tübingen & co. pursuing a Gaussian message propagation - much more practical.

Or, can we justify a relaxation of Bayesian PNM? “Cutting some of the dependencies”?

Turns out this is not a new problem at all...
A Conundrum

Is Bayesian PNM the unattainable dream?

fully Bayesian PNM pipeline in Oates, Cockayne and Aykroyd (2017)
vector-valued quadrature work of FX (in prep.)

Tübingen & co. pursuing a Gaussian message propagation - much more practical.

Or, can we justify a relaxation of Bayesian PNM? “Cutting some of the dependencies”?

Turns out this is not a new problem at all...
A Conundrum

Is Bayesian PNM the unattainable dream?

fully Bayesian PNM pipeline in Oates, Cockayne and Aykroyd (2017)
vector-valued quadrature work of FX (in prep.)

Tübingen & co. pursuing a Gaussian message propagation - much more practical.

Or, can we justify a relaxation of Bayesian PNM? “Cutting some of the dependencies”?

Turns out this is not a new problem at all...
Lawrence and I started wondering about these questions in the context of models of plankton population growth, back in 2012. Plankton growth is affected by ocean temperatures. These temperatures are not measured everywhere at all times, but one can first use a geophysics model to infer temperatures at desired locations and times. Then these estimated temperatures can be used as inputs (or “forcings”) to model plankton growth. We were wondering whether we should instead define a joint model of temperatures + plankton, to take the uncertainty of temperatures into account. Parslow, Cressie, Campbell, Jones & Murray (2013) provide an example of plankton model where temperatures are considered fixed, which is common practice.


Motivated by bias due to neglect.
Lawrence and I started wondering about these questions in the context of models of plankton population growth, back in 2012. Plankton growth is affected by ocean temperatures. These temperatures are not measured everywhere at all times, but one can first use a geophysics model to infer temperatures at desired locations and times. Then these estimated temperatures can be used as inputs (or “forcings”) to model plankton growth. We were wondering whether we should instead define a joint model of temperatures + plankton, to take the uncertainty of temperatures into account. Parslow, Cressie, Campbell, Jones & Murray (2013) provide an example of plankton model where temperatures are considered fixed, which is common practice.


Motivated by bias due to neglect.
Yesterday came out on arXiv a joint paper by Pierre Jacob, Lawrence Murray, Chris Holmes and myself, Better together? Statistical learning in models made of modules, paper that was conceived during the MCMski meeting in Chamonix, 2014! Indeed it is mostly due to Martyn Plummer’s talk at this meeting about the cut issue that we started to work on this topic at the fringes of the [standard] Bayesian world. Fringes because a standard Bayesian approach to the problem would always lead to use the entire dataset and the entire model to infer about a parameter of interest.

https://xianblog.wordpress.com/2017/08/31/better-together-2/

Motivated by reduction in computational cost.
Yesterday came out on arXiv a joint paper by Pierre Jacob, Lawrence Murray, Chris Holmes and myself, Better together? Statistical learning in models made of modules, paper that was conceived during the MCMski meeting in Chamonix, 2014! Indeed it is mostly due to Martyn Plummer’s talk at this meeting about the cut issue that we started to work on this topic at the fringes of the [standard] Bayesian world. Fringes because a standard Bayesian approach to the problem would always lead to use the entire dataset and the entire model to infer about a parameter of interest.

https://xianblog.wordpress.com/2017/08/31/better-together-2/

Motivated by reduction in computational cost.
In modern applications, statisticians are faced with integrating heterogeneous data modalities relevant for an inference, prediction, or decision problem. In such circumstances, it is convenient to use a graphical model to represent the statistical dependencies, via a set of connected modules, each relating to a specific data modality, and drawing on specific domain expertise in their development. In principle, given data, the conventional statistical update then allows for coherent uncertainty quantification and information propagation through and across the modules. However, misspecification of any module can contaminate the estimate and update of others, often in unpredictable ways. In various settings, particularly when certain modules are trusted more than others, practitioners have preferred to avoid learning with the full model in favor of approaches that restrict the information propagation between modules, for example by restricting propagation to only particular directions along the edges of the graph. In this article, we investigate why these modular approaches might be preferable to the full model in misspecified settings. We propose principled criteria to choose between modular and full-model approaches. The question arises in many applied settings, including large stochastic dynamical systems, meta-analysis, epidemiological models, air pollution models, pharmacokinetics-pharmacodynamics, and causal inference with propensity scores.

Motivated by bias due to misspecification.
In modern applications, statisticians are faced with integrating heterogeneous data modalities relevant for an inference, prediction, or decision problem. In such circumstances, it is convenient to use a graphical model to represent the statistical dependencies, via a set of connected modules, each relating to a specific data modality, and drawing on specific domain expertise in their development. In principle, given data, the conventional statistical update then allows for coherent uncertainty quantification and information propagation through and across the modules. However, misspecification of any module can contaminate the estimate and update of others, often in unpredictable ways. In various settings, particularly when certain modules are trusted more than others, practitioners have preferred to avoid learning with the full model in favor of approaches that restrict the information propagation between modules, for example by restricting propagation to only particular directions along the edges of the graph. In this article, we investigate why these modular approaches might be preferable to the full model in misspecified settings. We propose principled criteria to choose between modular and full-model approaches. The question arises in many applied settings, including large stochastic dynamical systems, meta-analysis, epidemiological models, air pollution models, pharmacokinetics-pharmacodynamics, and causal inference with propensity scores.


Motivated by bias due to misspecification.
Numerical Integration Thought Experiment

\[
\int_0^1 f(x) \, dx = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx
\]

Regardless of the prior, can we just ignore the loss of information across 1/2 and use the RHS?
Models Made of Modules
Thought Experiment on Two Variables

$\Theta_1, \Theta_2 =$ unknown parameters
$Y_1, Y_2 =$ observed data
top box = module 1, bottom box = module 2
Thought Experiment on Two Variables

$\Theta_1, \Theta_2 = \text{unknown parameters}$

$Y_1, Y_2 = \text{observed data}$

top box = module 1, bottom box = module 2
Candidate IPRs

- To fix ideas, suppose that our aim is to make inferences about the parameter $\Theta_1$.
- In principle we can use any, all or none of the data ($Y_1, Y_2$).
- From Zellner’s work, we know that the optimal IPR is Bayes’ rule:

$$p_{\text{Bayes}}(\theta_1 | Y_1, Y_2) = \frac{p(Y_1)}{p(Y_1, Y_2)} \frac{p(\theta_1 | Y_1)}{\text{posterior from module 1}} \times \frac{p(Y_2 | \theta_1)}{\text{feedback from module 2}}$$

- Another IPR could choose to “cut” the feedback from module 2:

$$p_{\text{cut 1}}(\theta_1 | Y_1, Y_2) := p(\theta_1 | Y_1)$$

or indeed $p_{\text{cut 2}} := p(\theta_1 | Y_2)$.
- Further, an IPR can choose to ignore the data entirely:

$$p_{\text{ignore}}(\theta_1 | Y_1, Y_2) := p(\theta_1)$$

Candidate IPRs

- To fix ideas, suppose that our aim is to make inferences about the parameter $\Theta_1$.
- In principle we can use any, all or none of the data $(Y_1, Y_2)$.
- From Zellner’s work, we know that the optimal IPR is Bayes’ rule:
  \[
  p_{\text{Bayes}}(\theta_1 | Y_1, Y_2) = \frac{p(Y_1)}{p(Y_1, Y_2)} \left( \frac{p(\theta_1 | Y_1)}{p(\theta_1 | Y_1)} \right) \times \frac{p(Y_2 | \theta_1)}{p(Y_2 | \theta_1)}
  \]
  posterior from module 1 feedback from module 2

- Another IPR could choose to “cut” the feedback from module 2:
  \[
  p_{\text{cut 1}}(\theta_1 | Y_1, Y_2) := \frac{p(\theta_1 | Y_1)}{p(\theta_1 | Y_1)}
  \]
  or indeed $p_{\text{cut 2}} := p(\theta_1 | Y_2)$.
- Further, an IPR can choose to ignore the data entirely:
  \[
  p_{\text{ignore}}(\theta_1 | Y_1, Y_2) := \frac{p(\theta_1)}{p(\theta_1)}
  \]
  prior from module 1

Candidate IPRs

- To fix ideas, suppose that our aim is to make inferences about the parameter \( \Theta_1 \).
- In principle we can use any, all or none of the data \((Y_1, Y_2)\).
- From Zellner’s work, we know that the optimal IPR is Bayes’ rule:

\[
p_{\text{Bayes}}(\theta_1 | Y_1, Y_2) = \frac{p(Y_1)}{p(Y_1, Y_2)} \frac{p(\theta_1 | Y_1)}{p(Y_1 | \theta_1)} \times \frac{p(Y_2 | \theta_1)}{p(Y_2 | \theta_1)}
\]

- posterior from module 1 feedback from module 2

- Another IPR could choose to “cut” the feedback from module 2:

\[
p_{\text{cut 1}}(\theta_1 | Y_1, Y_2) := \frac{p(\theta_1 | Y_1)}{p(\theta_1 | Y_1)}
\]

- or indeed \( p_{\text{cut 2}} := p(\theta_1 | Y_2) \).
- Further, an IPR can choose to ignore the data entirely:

\[
p_{\text{ignore}}(\theta_1 | Y_1, Y_2) := \frac{p(\theta_1)}{p(\theta_1)}
\]

- prior from module 1

Candidate IPRs

- To fix ideas, suppose that our aim is to make inferences about the parameter $\Theta_1$.
- In principle we can use any, all or none of the data ($Y_1, Y_2$).
- From Zellner’s work, we know that the optimal IPR is Bayes’ rule:

$$p_{\text{Bayes}}(\theta_1|Y_1, Y_2) = \frac{p(Y_1)}{p(Y_1, Y_2)} p(\theta_1|Y_1) \times p(Y_2|\theta_1)$$

  - posterior from module 1
  - feedback from module 2

- Another IPR could choose to “cut” the feedback from module 2:

$$p_{\text{cut 1}}(\theta_1|Y_1, Y_2) := p(\theta_1|Y_1)$$

  or indeed $p_{\text{cut 2}} := p(\theta_1|Y_2)$.

- Further, an IPR can choose to ignore the data entirely:

$$p_{\text{ignore}}(\theta_1|Y_1, Y_2) := p(\theta_1)$$

  or indeed $p_{\text{ignore}} := p(\theta_1)$.

Candidate IPRs

- To fix ideas, suppose that our aim is to make inferences about the parameter $\Theta_1$.
- In principle we can use any, all or none of the data $(Y_1, Y_2)$.
- From Zellner’s work, we know that the optimal IPR is Bayes’ rule:
  \[
  p_{Bayes}(\theta_1 | Y_1, Y_2) = \frac{p(Y_1)}{p(Y_1, Y_2)} p(\theta_1 | Y_1) \times \left( \frac{p(Y_2 | \theta_1)}{p(Y_2 | \theta_1)} \right)
  \]

  posterior from module 1 \hspace{1cm} feedback from module 2

- Another IPR could choose to “cut” the feedback from module 2:
  \[
  p_{cut\,1}(\theta_1 | Y_1, Y_2) := p(\theta_1 | Y_1)
  \]

  or indeed $p_{cut\,2} := p(\theta_1 | Y_2)$.
- Further, an IPR can choose to ignore the data entirely:
  \[
  p_{ignore}(\theta_1 | Y_1, Y_2) := p(\theta_1)
  \]

  prior from module 1

Candidate IPRs

- To fix ideas, suppose that our aim is to make inferences about the parameter $\Theta_1$.
- In principle we can use any, all or none of the data $(Y_1, Y_2)$.
- From Zellner’s work, we know that the optimal IPR is Bayes’ rule:

  \[
  p_{\text{Bayes}}(\theta_1 | Y_1, Y_2) = \frac{p(Y_1)}{p(Y_1, Y_2)} \times \frac{p(\theta_1 | Y_1)}{p(\theta_1 | Y_1)} \times \frac{p(Y_2 | \theta_1)}{p(Y_2 | \theta_1)}
  \]

  \[
  \text{posterior from module 1} \quad \times \quad \text{feedback from module 2}
  \]

- Another IPR could choose to “cut” the feedback from module 2:

  \[
  p_{\text{cut 1}}(\theta_1 | Y_1, Y_2) := \frac{p(\theta_1 | Y_1)}{p(\theta_1 | Y_1)}
  \]

  \[
  \text{posterior from module 1}
  \]

  or indeed $p_{\text{cut 2}} := p(\theta_1 | Y_2)$.

- Further, an IPR can choose to ignore the data entirely:

  \[
  p_{\text{ignore}}(\theta_1 | Y_1, Y_2) := \frac{p(\theta_1)}{p(\theta_1)}
  \]

  \[
  \text{prior from module 1}
  \]

Candidate IPRs

- To fix ideas, suppose that our aim is to make inferences about the parameter $\Theta_1$.
- In principle we can use any, all or none of the data ($Y_1, Y_2$).
- From Zellner's work, we know that the optimal IPR is Bayes' rule:

$$p_{\text{Bayes}}(\theta_1|Y_1, Y_2) = \frac{p(Y_1)}{p(Y_1, Y_2)} \frac{p(\theta_1|Y_1)}{p(Y_1, Y_2)} \times \frac{p(Y_2|\theta_1)}{p(Y_2|\theta_1)}$$

- Another IPR could choose to “cut” the feedback from module 2:

$$p_{\text{cut}_1}(\theta_1|Y_1, Y_2) := \frac{p(\theta_1|Y_1)}{p(Y_1, Y_2)}$$

or indeed $p_{\text{cut}_2} := p(\theta_1|Y_2)$.
- Further, an IPR can choose to ignore the data entirely:

$$p_{\text{ignore}}(\theta_1|Y_1, Y_2) := \frac{p(\theta_1)}{p(Y_1, Y_2)}$$

Aim of the Paper

Work in the so-called *M-open* framework:

Let $\pi$ denote the true data-generating distribution of $Y = (Y_1, Y_2)$. Then the statistical model (i.e. the likelihood) $p(Y|\theta)$ is called *misspecified* if:

$$\nexists \theta = (\theta_1, \theta_2) \text{ s.t. } \pi = p(Y|\theta).$$

Aim is to define a (computable) criterion to select between the belief distributions / IPRs

$$p_{\text{Bayes}}(\theta|Y), \ p_{\text{cut} \ 1}(\theta|Y), \ p_{\text{cut} \ 2}(\theta|Y), \ p_{\text{ignore}}(\theta|Y).$$

In general, this depends what the belief distribution will be used for...
Aim of the Paper

Work in the so-called *M-open* framework:

Let $\pi$ denote the true data-generating distribution of $Y = (Y_1, Y_2)$. Then the statistical model (i.e. the likelihood) $p(Y|\theta)$ is called *misspecified* if:

$$\nexists \theta = (\theta_1, \theta_2) \text{ s.t. } \pi = p(Y|\theta).$$

Aim is to define a (computable) criterion to select between the belief distributions / IPRs

$$p_{\text{Bayes}}(\theta|Y), \quad p_{\text{cut 1}}(\theta|Y), \quad p_{\text{cut 2}}(\theta|Y), \quad p_{\text{ignore}}(\theta|Y).$$

In general, this depends what the belief distribution will be used for...
Consider a general IPR $p_\bullet(\theta|Y)$.

Consider a utility function $u(\omega, \delta)$ which expresses the reward for taking the action $\delta$ when the state of the world is $\omega$, an unknown of interest.

Assume a known link function $\rho(\omega|\theta)$.

If we use an IPR to take an action, we would take the $M$-optimal act

$$d_\bullet(Y) = \arg \max_d \int u(\omega, d) \rho(\omega|\theta) p_\bullet(\theta|Y) d\omega d\theta.$$ 

In the real world, the expected utility of this IPR is

$$U_\bullet = \int u(\omega, d_\bullet(Y(\omega))) \pi(\omega) d\omega.$$ 

This suggests that we ought to pick the IPR $\bullet$ for which $U_\bullet$ is highest.

Two problems: (1) choice of $u$; (2) $U_\bullet$ depends on $\pi$, which is unknown.
Consider a general IPR $p_\bullet(\theta|Y)$.

Consider a utility function $u(\omega, \delta)$ which expresses the reward for taking the action $\delta$ when the state of the world is $\omega$, an unknown of interest.

Assume a known link function $\rho(\omega|\theta)$.

If we use an IPR to take an action, we would take the *M-optimal act*

$$d_\bullet(Y) = \arg \max_d \int u(\omega, d)\rho(\omega|\theta)p_\bullet(\theta|Y)d\omega d\theta.$$  

In the real world, the *expected utility* of this IPR is

$$U_\bullet = \int u(\omega, d_\bullet(Y(\omega)))\pi(\omega)d\omega.$$  

This suggests that we ought to pick the IPR $\bullet$ for which $U_\bullet$ is highest.

Two problems: (1) choice of $u$; (2) $U_\bullet$ depends on $\pi$, which is unknown.
Decision Theory for a Misspecified Model

Consider a general IPR \( p_\bullet(\theta|Y) \).

Consider a utility function \( u(\omega, \delta) \) which expresses the reward for taking the action \( \delta \) when the state of the world is \( \omega \), an unknown of interest.

Assume a known link function \( \rho(\omega|\theta) \).

If we use an IPR to take an action, we would take the \textit{M-optimal act}

\[
d_\bullet(Y) = \arg \max_d \int u(\omega, d) \rho(\omega|\theta) p_\bullet(\theta|Y) d\omega d\theta.
\]

In the real world, the \textit{expected utility} of this IPR is

\[
U_\bullet = \int u(\omega, d_\bullet(Y(\omega))) \pi(\omega) d\omega.
\]

This suggests that we ought to pick the IPR \( \bullet \) for which \( U_\bullet \) is highest.

Two problems: (1) choice of \( u \); (2) \( U_\bullet \) depends on \( \pi \), which is unknown.
Consider a general IPR $p(\theta|Y)$.

Consider a utility function $u(\omega, \delta)$ which expresses the reward for taking the action $\delta$ when the state of the world is $\omega$, an unknown of interest.

Assume a known link function $\rho(\omega|\theta)$.

If we use an IPR to take an action, we would take the $M$-optimal act

$$d(\omega) = \arg \max_d \int u(\omega, d) \rho(\omega|\theta) p(\theta|Y) d\omega d\theta.$$  

In the real world, the expected utility of this IPR is

$$U = \int u(\omega, d(\omega(Y(\omega)))) \pi(\omega) d\omega.$$  

This suggests that we ought to pick the IPR $\bullet$ for which $U$ is highest.

Two problems: (1) choice of $u$; (2) $U$ depends on $\pi$, which is unknown.
Consider a general IPR $p_\bullet(\theta|Y)$.

Consider a utility function $u(\omega, \delta)$ which expresses the reward for taking the action $\delta$ when the state of the world is $\omega$, an unknown of interest.

Assume a known link function $\rho(\omega|\theta)$.

If we use an IPR to take an action, we would take the $M$-optimal act

$$d_\bullet(Y) = \arg \max_d \int u(\omega, d) \rho(\omega|\theta) p_\bullet(\theta|Y) d\omega d\theta.$$ 

In the real world, the expected utility of this IPR is

$$U_\bullet = \int u(\omega, d_\bullet(Y(\omega))) \pi(\omega) d\omega.$$ 

This suggests that we ought to pick the IPR $\bullet$ for which $U_\bullet$ is highest.

Two problems: (1) choice of $u$; (2) $U_\bullet$ depends on $\pi$, which is unknown.
Decision Theory for a Misspecified Model

Consider a general IPR \( p_\bullet(\theta|Y) \).

Consider a utility function \( u(\omega, \delta) \) which expresses the reward for taking the action \( \delta \) when the state of the world is \( \omega \), an unknown of interest.

Assume a known link function \( \rho(\omega|\theta) \).

If we use an IPR to take an action, we would take the \textit{M-optimal act}

\[
d_\bullet(Y) = \arg \max_d \int u(\omega, d) \rho(\omega|\theta) p_\bullet(\theta|Y) d\omega d\theta.
\]

In the real world, the \textit{expected utility} of this IPR is

\[
U_\bullet = \int u(\omega, d_\bullet(Y(\omega))) \pi(\omega) d\omega.
\]

This suggests that we ought to pick the IPR \( \bullet \) for which \( U_\bullet \) is highest.

Two problems: (1) choice of \( u \); (2) \( U_\bullet \) depends on \( \pi \), which is unknown.
Let

- $\omega = Y^*$, a future observation to be predicted
- $d \in \mathcal{P}_Y$, a predictive distribution for the r.v. $Y^*$
- $\rho(\omega|\theta) = p(Y^*|\theta)$, link function is predictive likelihood

Then the default utility is the *logarithmic scoring rule*:

$$u(\omega, d) = \log d(Y^*)$$

i.e. the log-likelihood of $Y^*$ under the model $d(Y^*)$. 

Problem 1: A Default Utility

Let

- $\omega = Y^*$, a future observation to be predicted
- $d \in \mathcal{P}_Y$, a predictive distribution for the r.v. $Y^*$
- $\rho(\omega|\theta) = p(Y^*|\theta)$, link function is predictive likelihood

Then the default utility is the *logarithmic scoring rule*:

$$u(\omega, d) = \log d(Y^*)$$

i.e. the log-likelihood of $Y^*$ under the model $d(Y^*)$. 
Problem 2: An Approximation to $U_\bullet$

Require that the data $Y$ can be divided into independent data points $Y = (Y^1, \ldots, Y^n)$.

Then can proceed with some form of cross-validation; e.g.

$$U_\bullet = \int u(\omega, d_\bullet(Y(\omega)))\pi(\omega)d\omega$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \log \int p(Y^i|\theta)p_\bullet(\theta|Y^{(-i)})d\theta$$

So can approximate the expected utilities

$$U_{\text{Bayes}}, \quad U_{\text{cut 1}}, \quad U_{\text{cut 2}}, \quad U_{\text{ignore}}$$

to see which ought to be preferred.

If $U_{\text{Bayes}}$ is not largest, then we have a justification for not using the Bayesian IPR.
Problem 2: An Approximation to $U_\star$

Require that the data $Y$ can be divided into independent data points $Y = (Y^1, \ldots, Y^n)$.

Then can proceed with some form of cross-validation; e.g.

$$U_\star = \int u(\omega, d_\star(Y(\omega)))\pi(\omega)d\omega$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \log \int p(Y^i|\theta)p_\star(\theta|Y^{(-i)})d\theta$$

So can approximate the expected utilities

$$U_{\text{Bayes}}, \ U_{\text{cut 1}}, \ U_{\text{cut 2}}, \ U_{\text{ignore}}$$

to see which ought to be preferred.

If $U_{\text{Bayes}}$ is not largest, then we have a justification for not using the Bayesian IPR.
Problem 2: An Approximation to $U_\bullet$

Require that the data $Y$ can be divided into independent data points $Y = (Y^1, \ldots, Y^n)$.

Then can proceed with some form of cross-validation; e.g.

$$U_\bullet = \int u(\omega, d_\bullet(Y(\omega)))\pi(\omega)d\omega$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \log \int p(Y^i|\theta)p_\bullet(\theta|Y^{(-i)})d\theta$$

So can approximate the expected utilities $U_{\text{Bayes}}, U_{\text{cut 1}}, U_{\text{cut 2}}, U_{\text{ignore}}$ to see which ought to be preferred.

If $U_{\text{Bayes}}$ is not largest, then we have a justification for not using the Bayesian IPR.
Problem 2: An Approximation to $U_\bullet$

Require that the data $Y$ can be divided into independent data points $Y = (Y^1, \ldots, Y^n)$.

Then can proceed with some form of cross-validation; e.g.

$$U_\bullet = \int u(\omega, d_\bullet(Y(\omega)))\pi(\omega)d\omega$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \log \int p(Y^i|\theta)p_\bullet(\theta|Y^{(-i)})d\theta$$

So can approximate the expected utilities

$$U_{\text{Bayes}}, \quad U_{\text{cut 1}}, \quad U_{\text{cut 2}}, \quad U_{\text{ignore}}$$

to see which ought to be preferred.

If $U_{\text{Bayes}}$ is not largest, then we have a justification for not using the Bayesian IPR.
Illustration
Model Misspecification for $p(Y_2|\theta_1, \theta_2)$

Module 1:
- $Y_1 = (Y_1^1, \ldots, Y_{n1}^1)$, $Y_1^i \sim \mathcal{N}(\theta_1, 1)$ i.i.d., $n_1 = 100$
- $\theta_1 \sim \mathcal{N}(0, \lambda_1^{-1})$

Module 2:
- $Y_2 = (Y_2^1, \ldots, Y_{n2}^2)$, $Y_2^i \sim \mathcal{N}(\theta_1 + \theta_2, 1)$ i.i.d., $n_2 = 1000$
- $\theta_2 \sim \mathcal{N}(0, \lambda_2^{-1})$

Results for $\theta_1^{\text{true}} = 0$, $\theta_2^{\text{true}} = 1$, $\lambda_1 = 1$, $\lambda_2 = 100$, so module 2 is misspecified:
Model Misspecification for $p(Y_2|\theta_1, \theta_2)$

Module 1:
- $Y_1 = (Y_{1}^{1}, \ldots, Y_{1}^{n_1})$, $Y_1^i \sim \mathcal{N}(\theta_1, 1)$ i.i.d., $n_1 = 100$
- $\theta_1 \sim \mathcal{N}(0, \lambda_1^{-1})$

Module 2:
- $Y_2 = (Y_{2}^{1}, \ldots, Y_{2}^{n_2})$, $Y_2^i \sim \mathcal{N}(\theta_1 + \theta_2, 1)$ i.i.d., $n_2 = 1000$
- $\theta_2 \sim \mathcal{N}(0, \lambda_2^{-1})$

Results for $\theta_1^{\text{true}} = 0$, $\theta_2^{\text{true}} = 1$, $\lambda_1 = 1$, $\lambda_2 = 100$, so module 2 is misspecified:
Model Misspecification for $p(Y_2|\theta_1, \theta_2)$

Module 1:

- $Y_1 = (Y_1^1, \ldots, Y_1^{n_1})$, $Y_1^i \sim \mathcal{N}(\theta_1, 1)$ i.i.d., $n_1 = 100$
- $\theta_1 \sim \mathcal{N}(0, \lambda_1^{-1})$

Module 2:

- $Y_2 = (Y_2^1, \ldots, Y_2^{n_2})$, $Y_2^i \sim \mathcal{N}(\theta_1 + \theta_2, 1)$ i.i.d., $n_2 = 1000$
- $\theta_2 \sim \mathcal{N}(0, \lambda_2^{-1})$

Results for $\theta_1^{\text{true}} = 0$, $\theta_2^{\text{true}} = 1$, $\lambda_1 = 1$, $\lambda_2 = 100$, so module 2 is misspecified:
Model Misspecification for $p(Y_2|\theta_1, \theta_2)$

Module 1:
- $Y_1 = (Y_1^1, \ldots, Y_1^{n_1}), \ Y_1^i \sim \mathcal{N}(\theta_1, 1)$ i.i.d., $n_1 = 100$
- $\theta_1 \sim \mathcal{N}(0, \lambda_1^{-1})$

Module 2:
- $Y_2 = (Y_2^1, \ldots, Y_2^{n_2}), \ Y_2^i \sim \mathcal{N}(\theta_1 + \theta_2, 1)$ i.i.d., $n_2 = 1000$
- $\theta_2 \sim \mathcal{N}(0, \lambda_2^{-1})$

Results for $\theta_1^{\text{true}} = 0, \ \theta_2^{\text{true}} = 1, \ \lambda_1 = 1, \ \lambda_2 = 100$, so module 2 is misspecified:

![Density plots for parameters](image-url)
Model Misspecification for $p(Y_2|\theta_1, \theta_2)$

Module 1:
- $Y_1 = (Y_1^1, \ldots, Y_{n_1}^1)$, $Y_i^1 \sim \mathcal{N}(\theta_1, 1)$ i.i.d., $n_1 = 100$
- $\theta_1 \sim \mathcal{N}(0, \lambda_1^{-1})$

Module 2:
- $Y_2 = (Y_2^1, \ldots, Y_{n_2}^2)$, $Y_i^2 \sim \mathcal{N}(\theta_1 + \theta_2, 1)$ i.i.d., $n_2 = 1000$
- $\theta_2 \sim \mathcal{N}(0, \lambda_2^{-1})$

Results for $\theta_1^{\text{true}} = 0$, $\theta_2^{\text{true}} = 1$, $\lambda_1 = 1$, $\lambda_2 = 100$, so module 2 is misspecified:

<table>
<thead>
<tr>
<th></th>
<th>predictive score on $Y_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>module 1</td>
<td>-144.5</td>
</tr>
<tr>
<td>prior</td>
<td>-151.4</td>
</tr>
<tr>
<td>full model</td>
<td>-165</td>
</tr>
<tr>
<td>module 2</td>
<td>-188.8</td>
</tr>
</tbody>
</table>


The Remainder

*waves hands*

The remainder of the paper contains:

- Some “real world” examples that support the use of non-Bayesian IPR.
- Discussion on non-Bayesian IPR for confidentiality.
- Discussion on how to sample from a “cut” belief distribution, which is just about being careful.
Summary
Main points:

- Decision-theoretic considerations can provide philosophical grounds for not using the Bayesian IPR.

- Moreover, such a framework may be useful to assess model misspecification (which could be particularly acute for PNM).

- Unclear about how this scales to large work-flows, or whether it is computationally practical (as $U_\star$ is a marginal likelihood).

Other promising directions include approximate Bayesian PNM, but this needs further work.

Something to think about for April...
Main points:

▶ Decision-theoretic considerations can provide philosophical grounds for not using the Bayesian IPR.

▶ Moreover, such a framework may be useful to assess model misspecification (which could be particularly acute for PNM).

▶ Unclear about how this scales to large work-flows, or whether it is computationally practical (as $U_\bullet$ is a marginal likelihood).

Other promising directions include approximate Bayesian PNM, but this needs further work.

Something to think about for April...
The SAMSI working group on probabilistic numerics is meeting for a workshop in April:

https://prob-num.github.io/

Places are strictly limited (by our catering budget) - registration details to follow by email.

Please remember to acknowledge SAMSI in publications that benefit from this WG!

Lastly, interested speakers for this series should please contact FX.