Why Gelman “hates” Bayesian model comparison

Richard E. Turner (ret26@cam.ac.uk)
The debate

- **Gelman:** A lot has been written in statistics about “parsimony” - that is, the desire to explain phenomena using fewer parameters - but I’ve never seen any good general justification for parsimony.
The debate

- **Gelman:** A lot has been written in statistics about “parsimony” - that is, the desire to explain phenomena using fewer parameters - but I’ve never seen any good general justification for parsimony.

- **Radford:** Sometimes a simple model will outperform a more complex model ... Nevertheless, I believe that deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well.
The debate

- **Gelman:** A lot has been written in statistics about “parsimony” - that is, the desire to explain phenomena using fewer parameters - but I’ve never seen any good general justification for parsimony.

- **Radford:** Sometimes a simple model will outperform a more complex model ... Nevertheless, I believe that deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well.

- **Gelman:** Exactly! I hate all the Occam-factor stuff that MacKay talks about.
The debate

• **Gelman**: A lot has been written in statistics about “parsimony” - that is, the desire to explain phenomena using fewer parameters - but I’ve never seen any good general justification for parsimony.

• **Radford**: Sometimes a simple model will outperform a more complex model ... Nevertheless, I believe that deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well.

• **Gelman**: Exactly! I hate all the Occam-factor stuff that MacKay talks about.

• **Mackay**: When you said you disagree with me on Occam factors I think what you meant was that you agree with me on them...
Gelman, Radford and Mackay agree on:

1. “We should be using models the size of a house”

2. Cranking the handle of Bayesian inference “embodies Occam's razor”

But Gelman is right, there is a:

3. practical difficulty with Bayesian model comparison (even when computation is exact).
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H) \]
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H) \]
\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H)^N \overset{N \to \infty}{\Rightarrow} p(D|\mu, H) \]
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H) \xrightarrow{N \to \infty} p(D|\mu, H) \]  

centre/width

indep. of prior
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H)^N \xrightarrow{N \to \infty} p(D|\mu, H) \]

\[ p(D|H) = \int d\mu \ p(\mu|H)p(D|\mu, H) \]

centre/width
indep. of prior
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H)^{N \rightarrow \infty} \Rightarrow p(D|\mu, H) \]

indep. of prior

\[ p(D|H) = \int d\mu \, p(\mu|H)p(D|\mu, H) \]
Main idea

\[
p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H)^N \approx p(D|\mu, H) \quad \text{centre/width}
\]

\[
p(D|H) = \int d\mu \ p(\mu|H)p(D|\mu, H)
\]

\[
\approx p(D|\mu^*, H) \frac{\sigma_{\mu|D}}{\sigma_\mu}
\quad \text{(Laplace's approximation)}
\]
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H)^N \Rightarrow \infty \quad p(D|\mu, H) \]

\[ p(D|H) = \int d\mu \quad p(\mu|H)p(D|\mu, H) \]

\[ \approx p(D|\mu^*, H) \frac{\sigma_{\mu|D}}{\sigma_{\mu}} \]

Occam factor
penalises likelihood

\[ p(D|\mu^*, H) \]

\[ p(D|\mu, H) \]

\[ p(\mu|H) \]
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H) \overset{N \to \infty}{\Rightarrow} p(D|\mu, H) \]

\[ p(D|H) = \int d\mu \, p(\mu|H)p(D|\mu, H) \]

\[ \approx p(D|\mu^*, H) \]

Occam factor
penalises likelihood
posterior volume/prior volume

Centre/width
indep. of prior
Main idea

\[ p(\mu|D, H) \propto p(\mu|H)p(D|\mu, H)^N \Rightarrow p(D|\mu, H) \]

\[ p(D|H) = \int d\mu \ p(\mu|H)p(D|\mu, H) \]

\[ \approx p(D|\mu^*, H) \frac{\sigma_{\mu|D}}{\sigma_\mu} \]

Occam factor

penalises likelihood
posterior volume/prior volume
always depends on prior
indep. of prior
Simple example

- Posterior will not depend strongly on prior
- Marginal likelihood depends strongly on prior

\[ p(\mu|H_1) = \text{Norm}(\mu; 0, \sigma_\mu^2) \]
\[ p(x_n|\mu, H_1) = \text{Norm}(x_n; \mu, 1) \]

(Not sure it’s a great example, but simple to understand.)
Simple example
Simple example
Simple example
Simple example

\[ \text{posterior mean} \]

\[ \log p(D|H_i) \]
Simple example
Simple example

\[ \sigma, \mu, \text{posterior mean} \]

\[ \log p(D|H_i) \]
Simple example
Simple example
Simple example

\( p(\mu|D, H_1) \)

\( p(x|D, H_1) \)

stable
Simple example

$p(\mu|D, H_1)$
$p(x|D, H_1)$
stable

$p(D|H_1)$
sensitive
Simple example

easy to be here in complex models

\[ p(\mu | D, H_1) \]
\[ p(x | D, H_1) \]

stable

\[ p(D | H_1) \]
sensitive
Conclusions

• Discrete Bayesian model comparison: **beware the prior**
  – Uninformative priors dangerous (improper priors apocalyptic)
  – Perform a sensitivity analysis
  – Common tactic: convert model comparison into parameter estimation problem
  – Philosophical inconsistency - model comparison is just (discrete) inference

• **Posterior predictive tests**: can tell you in what way your model is wrong without needing another to compare to another model

Read both books...

www.inference.phy.cam.ac.uk

andrewgelman.com